Overlapping Schwarz methods with GenEO coarse spaces for indefinite and non-self-adjoint problems

Niall Bootland[*], Victorita Dolean[*,†], Ivan G. Graham[‡], Chupeng Ma[§], Robert Scheichl[‡,§]

Abstract

GenEO ('Generalised Eigenvalue problems on the Overlap') is a method for computing an operator-dependent spectral coarse space to be combined with local solves on subdomains to form a robust parallel domain decomposition preconditioner for elliptic PDEs. It has previously been proved, in the self-adjoint and positive-definite case, that this method, when used as a preconditioner for conjugate gradients, yields iteration numbers which are completely independent of the heterogeneity of the coefficient field of the partial differential operator. We extend this theory to the case of convection–diffusion–reaction problems, which may be non-self-adjoint and indefinite, and whose discretisations are solved with preconditioned GMRES. The GenEO coarse space is defined here using a generalised eigenvalue problem based on a self-adjoint and positive-definite subproblem. We prove estimates on GMRES iteration counts which are independent of the variation of the coefficient of the diffusion term in the operator and depend only very mildly on variations of the other coefficients. These are proved under the assumption that the subdomain diameter is sufficiently small and the eigenvalue tolerance for building the coarse space is sufficiently large. While the iteration number estimates do grow as the non-self-adjointness and indefiniteness of the operator increases, practical tests indicate the deterioration is much milder. Thus we obtain an iterative solver which is efficient in parallel and very effective for a wide range of convection–diffusion–reaction problems.

MSC Classes: `65N22, 65N55, 65F10`

Keywords: elliptic PDEs, finite element methods, domain decomposition, preconditioning, GMRES, convection–diffusion–reaction, spectral coarse space, robustness

1 Introduction

This paper is concerned with the theory and implementation of two-level domain decomposition preconditioners for finite element systems arising from indefinite and/or non-self-adjoint elliptic boundary value problems with rough or highly variable coefficients. We analyse and test a method which is scalable with respect to the number of subdomains and also robust to heterogeneity, backed up by supporting theory. The method is developed for the following scalar second-order (convection–diffusion–reaction) PDE with homogeneous Dirichlet boundary conditions:

$$-\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu = f \qquad \text{in } \Omega, \tag{1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{1.1b}$$

[*]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, G1 1XH, UK; `niall.bootland@strath.ac.uk, work@victoritadolean.com`

[†]Laboratoire J.A. Dieudonné, CNRS, University Côte d'Azur, Nice, 06108, France

[‡]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK; `i.g.graham@bath.ac.uk`

[§]Institute for Applied Mathematics & Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany; `chupeng.ma@uni-heidelberg.de, r.scheichl@uni-heidelberg.de`

This is a peer reviewed, accepted author manuscript of the following research article: Bootland, N., Dolean, V., Graham, I. G., Ma, C., & Scheichl, R. (2022). Overlapping Schwarz methods with GenEO coarse spaces for indefinite and nonself-adjoint problems. IMA Journal of Numerical Analysis. https://doi.org/10.1093/imanum/drac036

posed on a bounded polygonal or Lipschitz polyhedral domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$. As well as a weak regularity assumption on the coefficients $A, \mathbf{b}$ and c (Assumption 2.1), we will assume that (1.1) has a unique weak solution $u \in H^1(\Omega)$ for all $f \in L^2(\Omega)$.

In overlapping domain decomposition methods, the global domain Ω is covered by a set of overlapping subdomains Ω_i, $i = 1, \dots N$, and the classical one-level additive Schwarz preconditioner is built from partial solutions on each subdomain. Since this preconditioner is not in general scalable as the number of subdomains grows, an additional global coarse solve is usually added to enhance scalability, as well as robustness with respect to coefficient heterogeneity.

Aim of the paper and related literature. The construction of suitable coarse spaces has enjoyed intensive recent interest. Since classical coarse spaces (piecewise polynomials on a coarse grid, e.g., [9, 38]) are not robust for highly heterogeneous problems, a number of groups have developed operator-dependent coarse spaces which are better adapted to the heterogeneity, an early example being [21]. Most recently, attention has focussed on 'spectral' coarse spaces which are built from selected modes of local generalised eigenvalue problems. The first proposals of this technique for domain decomposition preconditioning were [16, 17]. The use of local spectral information is important in the related but different context of robust approximation techniques for multiscale PDE problems, e.g., [2, 13, 29, 30].

This paper focuses on the GenEO coarse space, first proposed in [37] for general coercive, self-adjoint systems of PDEs (e.g., (1.1) with $\mathbf{b} = 0$ and $c \geq 0$). In [37], it was rigorously proved that (when combined with a one-level method and using preconditioned conjugate gradients as the iterative solver), the resulting algorithm enjoys not only scalability with respect to the number of subdomains, but also robustness with respect to coefficient variation. Recent contributions in this very active field include [1, 25, 3] and [36], while more complete lists of contributions can be found in the literature surveys in, e.g., in the introductions to [37] or [15].

While all the methods cited so far were developed for SPD problems, where the behaviour of the local eigenspaces can be more readily analysed, there has also been considerable (so far mostly empirical) progress on applying related techniques to non-coercive problems. Most of this has been for the Helmholtz problem at high frequency, although [31] contains rigorous recent work on saddle-point systems. For Helmholtz, the first publications include [11, 4], while recent work in [6] compares the performance of several spectral coarse spaces for high-frequency Helmholtz problems in a high performance environment. One of the most encouraging coarse spaces for Helmholtz problems (called 'H-GenEO' in [6, 5]) has promising robustness properties at high frequency, but, since it involves solving for local generalised eigenvalues of the Helmholtz operator, it is beyond the reach of the present theory. However, the fact that H-GenEO behaves similarly to the method analysed in this paper at lower frequencies, see [5], motivated the current work.

Our other primary motivation was to find out how well the GenEO technology works (both in theory and in practice) for general non-self-adjoint and indefinite PDEs of the form (1.1). The GenEO coarse space that we analyse here is the same as the space previously proposed and analysed in the SPD case. Thus our numerical results here indicate the results which can be expected when any standard GenEO code is used for the more general problem (1.1). The main contribution of the present paper therefore is the analysis of how GenEO performs as a preconditioner for GMRES when applied to (1.1) and, in particular, its dependence on heterogeneity, indefiniteness, and lack of self-adjointness.

Since the FE systems that arise from (1.1) are in general non-Hermitian, we use GMRES as the iterative solver and our main tool for convergence analysis is the 'Elman theory' ([14] or [38, Lemma C.11]), which requires an upper bound for the norm of the preconditioned matrix and a lower bound on the distance of its field of values from the origin. The number of GMRES iterates to achieve a given error tolerance is then a function of these two bounds. Thus, the dependence of these bounds on the parameters of the problem (e.g., mesh size, subdomain size, overlap, heterogeneity, and strength of indefiniteness/non-self-adjointness) is of paramount interest when analysing the robustness and scalability of the preconditioner. The use of the Elman theory in domain decomposition goes back to Cai and Widlund [9] (see also [38, Section 11]), who used classical linear piecewise polynomials on a coarse grid without attention to coefficient heterogeneity. Analogous results for Maxwell's equations were obtained in [20].

The main results of this paper. Our main theoretical result, Theorem 4.1, provides rigorous upper bounds on the coarse mesh diameter H and on the 'eigenvalue tolerance' Θ (for the local GenEO eigenproblems) which ensure that GMRES enjoys robust and mesh-independent convergence when applied to the preconditioned problem. Reducing H and Θ still further would improve GMRES convergence but increase cost: smaller Θ means that more local eigenfunctions are incorporated into the coarse space; smaller H means that more local eigenvalue problems have to be solved, although the size of the local subdomains becomes smaller and, in practice, fewer eigenfunctions are needed on each subdomain. Optimal choices of Θ and H are typically determined empirically. Since Theorem 4.1 is quite technical, by way of introduction we present here the result when it is specialised to the following simpler, yet instructive, example.

Example 1.1 (An indefinite but self-adjoint problem)**.** Consider problem (1.1) with $\mathbf{b} = \mathbf{0}$, and $c(x) = -\kappa$ where κ is a positive constant. That is

$$-\mathrm{div}(A\nabla u) - \kappa u = f, \tag{1.2}$$

with homogeneous Dirichlet boundary conditions. Let $a_{\min}$ and $a_{\max}$ be, respectively, lower and upper spectral bounds on A (see (2.3)). Without loss of generality, we assume that $a_{\min} = 1$ and that the diameter D_Ω of Ω satisfies $D_\Omega \leq 1$. (Otherwise the problem can be rescaled to achieve these requirements.) The GenEO coarse space is based on the m_i (dominant) eigenfunctions corresponding to the smallest eigenvalues $\lambda_1^i \leq \lambda_2^i \leq \ldots \leq \lambda_{m_i}^i$ of the generalised eigenvalue problem on Ω_i defined in (2.31). To obtain a robust rate of convergence for GMRES that depends only on k_0 (the maximum number of times any point is overlapped by the subdomains Ω_i), Theorem 4.1 proves that the following conditions on the 'eigenvalue tolerance' $\Theta := 1/\min_{i=1}^{N} \lambda_{m_i+1}^i$ and on the (coarse) mesh diameter H are sufficient:

$$H \ \lesssim \ \kappa^{-1} \quad \text{and} \quad \Theta^{1/2} \lesssim (C^*_{\mathrm{stab}} + 1)^{-1}\, \kappa^{-2}. \tag{1.3}$$

Here $C^*_{\mathrm{stab}} > 0$ is the stability constant for the adjoint problem to (1.1) and is defined in (2.10). It blows up as the distance of κ from an eigenvalue of the Dirichlet problem for $-\mathrm{div}(A\nabla u)$ decreases. The hidden constants in (1.3) depend only on k_0. Thus, under conditions (1.3) the rate of convergence of GMRES depends only on k_0. For constant A, the first condition in (1.3) is exactly the one required to ensure that the piecewise linear finite element discretization of (1.1) on a coarse mesh of diameter H is uniquely solvable (see, e.g., [9]).

The full statement of Theorem 4.1 for a spatially- and sign-varying c and for a nonzero, spatially varying convection coefficient $\mathbf{b}$ is structurally similar. Sufficient conditions for robustness are obtained by replacing κ^{-1} and κ^{-2} in the two bounds (1.3), respectively, by

$$(1 + \|\mathbf{b}\|_\infty + \|c\|_\infty)^{-1},$$

and

$$(1 + \|\mathbf{b}\|_\infty + \|c\|_\infty)^{-1}(1 + \|\mathbf{b}\|_\infty + \|\nabla\cdot\mathbf{b}\|_\infty + \|c\|_\infty)^{-1}.$$

By making a small change to the GenEO construction, the contribution from c in these bounds can be removed if c is uniformly non-negative. We remark also that our results do not depend on the spatial variation of A or c and depend on the variation of $\mathbf{b}$ only through its divergence.

The proof of Theorem 4.1 is obtained by estimating the norm and the field of values of the preconditioned problem in the energy inner product induced by the second order operator in

(1.1), and then applying the Elman estimate. Several technical difficulties have to be overcome, particularly: (i) since the regularity assumptions on the coefficients A, $\mathbf{b}$, and c are minimal, there is no a priori regularity for solutions of (1.1) beyond H^1; (ii) because of the indefiniteness and non-self-adjointness of (1.1), the existence and stability of solutions of both local and coarse space subproblems has to be established, and this is the first source of conditions on H and Θ; (iii) all estimates should be explicit in the coefficients A, $\mathbf{b}$, and c, and as sharp as possible.

After the proof of Theorem 4.1, we remark that restricting the result to the special case $\mathbf{b} = \mathbf{0}$ and $c = 0$, we obtain essentially the same result as in [37], which is interesting because the latter makes use of 'self-adjoint' technology, working with eigenvalue and condition number estimates and does not estimate the field of values as is required in [14].

In Section 5 we give substantial numerical experiments illustrating the theory. All experiments are on rectangular domains with uniform fine meshes. Highlights of these results include: (i) for problem (1.1), with $A = I$, $\mathbf{b} = 0$, and $c = -\kappa$ constant, the standard GenEO preconditioner is remarkably effective even up to about $\kappa = 1000$; (ii) in the case $A = I$, $c = 0$, and $\mathbf{b}$ variable, the method is effective up to about $\|\mathbf{b}\|_\infty = 1000$, with the results only degrading moderately as the divergence of $\mathbf{b}$ increases from zero; (iii) the results above hardly change when $A = I$ is changed to a highly heterogeneous field; (iv) provided the ratio of overlap width to subdomain diameter is held fixed as $h \to 0$, the dimension of the GenEO coarse space is independent of h; (v) the method provides a very effective preconditioner for systems arising from implicit time-stepping applied to convection–diffusion–reaction problems, and the preconditioner can be assembled once and used for all choices of (variable) time-steps.

We remark that, while Theorem 4.1 provides convergence estimates for 'weighted' GMRES applied in the 'energy' inner product, our numerical experiments use standard GMRES and the Euclidean inner product. In Corollary 4.4 we show that standard GMRES attains the same accuracy as weighted GMRES after a few extra iterations which grow only logarithmically in h^{-1} and in the contrast of the matrix A. The empirical observation that weighted and standard GMRES yield similar iteration counts in domain decomposition methods has been made before, e.g., [10, 22] and a result related but not identical to Corollary 4.4 is given in [19, Corollary 5.8].

In our experiments the rather stringent looking condition on the eigenvalue tolerance in (1.3) appears to be overly pessimistic and the coarse space sizes necessary to achieve full robustness in practice are moderate. Similar observations were made in [9], for homogeneous but indefinite problems and classical coarse spaces.

Plan of the paper. In Section 2 we give some background to domain decomposition for (1.1) and introduce the GenEO preconditioner. In Section 3 we prove some of the central technical estimates concerning solutions of the subproblems, which are then used in Section 4 to prove the main theorem. Finally, numerical experiments are detailed in Section 5.

2 Background Material

2.1 Problem formulation and discretization

The weak formulation of (1.1) is to find $u \in H_0^1(\Omega)$ such that

$$b(u, v) = (f, v), \quad \forall v \in H_0^1(\Omega), \tag{2.1}$$

where $f \in L^2(\Omega)$ and the bilinear form $b(\cdot,\cdot)\colon H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is defined by

$$b(u, v) = \int_\Omega \big(A\nabla u \cdot \nabla v + \mathbf{b} \cdot \nabla u\, v + cuv\big) dx. \tag{2.2}$$

Throughout the paper we make the following assumptions on the problem:

Assumption 2.1. *The coefficients (A, $\mathbf{b}$, c) and the right-hand side f are assumed to satisfy* *(i)* $A : \Omega \to \mathbb{R}^{d\times d}$ *is symmetric and there exists* $0 < a_{\min} < a_{\max}$ *such that*

$$a_{\min}|\xi|^2 \le A(x)\xi\cdot\xi \le a_{\max}|\xi|^2, \quad \textit{for all } x\in\Omega,\ \ \xi\in\mathbb{R}^d; \tag{2.3}$$

(ii) $\mathbf{b}\in (L^\infty(\Omega))^d$, $\nabla\cdot\mathbf{b}\in L^\infty(\Omega)$, $c\in L^\infty(\Omega)$, $f\in L^2(\Omega)$.

(iii) Without loss of generality, we assume that $a_{\min} = 1$ *and that the diameter* D_Ω *of the domain* Ω *satisfies* $D_\Omega \le 1$*; otherwise the problem can simply be rescaled to achieve this.*

(iv) In order to distinguish between strongly definite and strongly indefinite problems we allow a splitting of c*:*

$$c = c^+ + c^-, \quad \textit{with} \quad c^+(x)\ge 0 \quad \textit{for all } x\in\Omega, \quad \textit{and} \quad c^- := c - c^+. \tag{2.4}$$

A natural choice would be to choose c^+ to be the non-negative part of c, i.e.,

$$c^+(x) = \begin{cases} c(x) & \text{when } c(x)\ge 0,\\ 0 & \text{otherwise.}\end{cases}$$

With this choice, $c^- = 0$ if c is non-negative. However other splittings are possible, e.g., $c^+ = 0$, and $c^- = c$.

Our analysis treats (2.2) as a perturbation of the coercive bilinear form

$$a(u,v) = \int_\Omega \left(A\nabla u\cdot\nabla v + c^+ uv\right)\, dx, \quad u,v\in H_0^1(\Omega). \tag{2.5}$$

Using this we can write

$$b(u,v) = a(u,v) + (\mathbf{b}\cdot\nabla u, v) + (c^- u, v), \quad \text{for all} \quad u,v\in H_0^1(\Omega). \tag{2.6}$$

Alternatively, after integration by parts, we can rewrite $b(u,v)$ as

$$b(u,v) = a(u,v) - (\mathbf{b}\cdot\nabla v, u) + (\tilde{c}u, v), \quad \text{for all} \quad u,v\in H_0^1(\Omega), \tag{2.7}$$

with

$$\tilde{c} = c^- - \nabla\cdot\mathbf{b}, \tag{2.8}$$

where in (2.8) we make use of Assumption 2.1(ii). Both expressions for b will be useful later.

Notation 2.2. *For any subdomain* $\Omega'\subset\Omega$*, we write* $(\cdot,\cdot)_{\Omega'}$ *to denote the* $L^2(\Omega')$ *inner product, with induced norm* $\|\cdot\|_{L^2(\Omega')}$*. When* $\Omega' = \Omega$ *we write these more simply as* $(\cdot,\cdot)$ *and* $\|\cdot\|$ *respectively. Further, the* L^∞ *norm on* Ω *is denoted* $\|\cdot\|_\infty$*. For all other norms we indicate the space explicitly as a subscript, e.g.,* $\|\cdot\|_{H_0^1(\Omega)}$*. We will make extensive use of the energy norm*

$$\|u\|_a := \sqrt{a(u,u)}, \quad \text{for all } u\in H_0^1(\Omega). \tag{2.9}$$

Furthermore, we assume unique solvability for (2.1) and its adjoint:

Assumption 2.3. *(i) For any* $f\in L^2(\Omega)$*, we assume the problem* (2.1) *has a unique solution* $u\in H_0^1(\Omega)$ *and there exists a constant* $C_{\text{stab}} > 0$ *such that*

$$\|u\|_a \le C_{\text{stab}}\|f\|, \quad \textit{for all } f\in L^2(\Omega).$$

(ii) For the adjoint problem: seek $w\in H_0^1(\Omega)$ *such that* $b(v,w) = (f,v)$, $v\in H_0^1(\Omega)$*, we assume that, for any* $f\in L^2(\Omega)$*, there exists a unique solution* $w\in H_0^1(\Omega)$ *and a constant* $C^*_{\text{stab}} > 0$ *such that*

$$\|w\|_a \le C^*_{\text{stab}}\|f\|, \quad \textit{for all } f\in L^2(\Omega). \tag{2.10}$$

Note that, in contrast to other works (e.g., [9], [7, Chapter 5]), we make no *a priori* regularity assumptions on u beyond those implied by Assumptions 2.1 and 2.3. Thus, our only explicit assumption on A is its uniform positive-definiteness (2.3).

Let $\mathcal{T}_h$ be a shape-regular triangulation of the domain Ω consisting of triangles (2-d) or tetrahedra (3-d) of maximal diameter h. Let $V^h \subset H_0^1(\Omega)$ be any conforming finite element (FE) space. The Galerkin approximation of (2.1) seeks $u_h \in V^h$ such that

$$b(u_h, v) = (f, v), \quad \text{for all} \quad v \in V^h. \tag{2.11}$$

Let n denote the dimension of V^h and let $\{\phi_i\}_{i=1}^n$ be a basis for V^h. The linear system corresponding to (2.11) is given by

$$\mathbf{B}\mathbf{u} = \mathbf{f}, \tag{2.12}$$

where the matrix $\mathbf{B}$ is defined by $\mathbf{B}_{ij} = b(\phi_i, \phi_j)$ and $\mathbf{f} = (f_i) \in \mathbb{R}^n$ with $f_i = (f, \phi_i)$. The solvability of (2.11) is guaranteed by the following result from [34, Theorem 2].

Lemma 2.4 (Schatz & Wang, 1996)**.** *Suppose Assumptions 2.1 and 2.3 hold. Then there exists $h_0 > 0$, such that for each $0 < h < h_0$, (2.11) has a unique solution u_h. In addition, for any given $\varepsilon > 0$, there exists an $h_1 = h_1(\varepsilon)$, such that*

$$\|u - u_h\|_{H^1(\Omega)} \leq \varepsilon \|f\|, \tag{2.13}$$

for $0 < h < h_1$, where u is the solution of (2.1). The same holds true also for the FE solution to the adjoint problem, introduced in Assumption 2.3(ii).

Note that Lemma 2.4 implies that, for any $\varepsilon > 0$, there exists $h_2 := h_1(\varepsilon/(\sqrt{a_{\max} + \|c^+\|_\infty}))$ such that

$$\|u - u_h\|_a \leq \varepsilon \|f\|, \quad \text{for} \quad 0 < h < h_2. \tag{2.14}$$

To finish this subsection, we introduce the following Friedrichs inequality with an explicit constant for subsequent use.

Lemma 2.5 (Friedrichs inequality [28, Theorem 13.19])**.** *Let $\Omega' \subset \mathbb{R}^d$ be an open set that lies between two parallel hyperplanes. Then, for all $u \in H_0^1(\Omega')$,*

$$\|u\|_{L^2(\Omega')} \leq \frac{L}{\sqrt{2}} \|\nabla u\|_{L^2(\Omega')}, \tag{2.15}$$

where L is the distance between the two hyperplanes.

Combining (2.15) with Assumption 2.1(iii), for any subdomain $\Omega' \subset \Omega$ with diameter H, we have the estimate

$$\|u\|_{L^2(\Omega')} \leq \frac{H}{\sqrt{2}} \|\nabla u\|_{L^2(\Omega')} \leq \frac{H}{\sqrt{2}} \|u\|_{a,\Omega'}, \quad \text{for all} \quad u \in H_0^1(\Omega'). \tag{2.16}$$

2.2 Domain decomposition

In order to construct a one-level Schwarz preconditioner for (2.12), we first partition the domain Ω into a set of non-overlapping subdomains $\{\Omega_i'\}_{i=1}^N$. We assume the boundaries of the Ω_i' are resolved by $\mathcal{T}_h$. Each subdomain Ω_i' is extended to a domain Ω_i by adding one or more layers of adjoining mesh elements, thus creating an overlapping decomposition $\{\Omega_i\}_{i=1}^N$ of Ω. We assume that each Ω_i has a diameter H_i and set $H = \max_i\{H_i\}$.

For each $i = 1, \ldots, N$ we define

$$\widetilde{V}_i = \{v|_{\Omega_i} \; : \; v \in V^h\} \subset H^1(\Omega_i), \quad \text{and} \quad V_i = \{v \in \widetilde{V}_i \; : \; v|_{\partial\Omega_i} = 0\} \subset H_0^1(\Omega_i). \tag{2.17}$$

For any $u,\, v \in \widetilde{V}_i$ and $i = 1, \ldots, N$ we define the local bilinear forms

$$
\begin{aligned}
a_{\Omega_i}(u,v) &:= \int_{\Omega_i} \big(A\nabla u \cdot \nabla v \, + c^+ uv\big)\, dx, \\
b_{\Omega_i}(u,v) &:= \int_{\Omega_i} \big(A\nabla u \cdot \nabla v + \mathbf{b}\cdot \nabla u\, v + cuv\big) dx \\
&= a_{\Omega_i}(u,v) + (\mathbf{b}\cdot \nabla u, v)_{\Omega_i} + (c^-\, u, v)_{\Omega_i}.
\end{aligned} \tag{2.18}
$$

Then, in analogy to (2.9), we write

$$
\|u\|_{a,\Omega_i} = \sqrt{a_{\Omega_i}(u,u)}, \quad \text{for all } \ u \in V_i, \tag{2.19}
$$

which is a norm on V_i.

For any $v_i \in V_i$, let $R_i^\top v_i$ denote its zero extension to all of Ω. Then, because of (2.17),

$$
R_i^\top : V_i \to V^h, \quad i = 1, \ldots, N. \tag{2.20}
$$

The $L^2(\Omega)$ adjoint of $R_i^\top$ is the *restriction operator* $R_i \colon V^h \to V_i$, defined by $R_i v = v|_{\Omega_i}$. Using the extension operators, the bilinear forms in (2.18) can be defined equivalently by

$$
b_{\Omega_i}(u,v) := b(R_i^\top u, R_i^\top v), \quad a_{\Omega_i}(u,v) := a(R_i^\top u, R_i^\top v), \ \ \text{and} \ \ (u,v)_{\Omega_i} := (R_i^\top u, R_i^\top v), \tag{2.21}
$$

for all $u, v \in V_i$.

In matrix form, the classical one-level additive Schwarz preconditioner for (2.12) can now be defined as

$$
\mathbf{M}_{AS,1}^{-1} = \sum_{i=1}^{N} \mathbf{R}_i^T \mathbf{B}_i^{-1} \mathbf{R}_i, \quad \text{where} \quad \mathbf{B}_i := \mathbf{R}_i \mathbf{B} \mathbf{R}_i^T, \quad i = 1, \ldots, N. \tag{2.22}
$$

Here, $\mathbf{R}_i$ denotes the matrix representation of R_i with respect to the basis $\{\phi_i\}_{i=1}^n$ of V^h.

It is well-known that, since each application of $\mathbf{M}_{AS,1}^{-1}$ exchanges information only between neighbouring subdomains, this preconditioner is generally not scalable with respect to the number of subdomains N. To facilitate a global exchange of information, a solve in a coarse space is usually added. Let $V_0 \subset V^h$ be such a coarse space; a classical example is an $H_0^1(\Omega)$-conforming FE space corresponding to a coarse triangulation $\mathcal{T}_0$ of Ω, chosen so that $\mathcal{T}_h$ is a (regular) refinement of $\mathcal{T}_0$. Let $R_0^\top : V_0 \to V^h$ denote the natural embedding (i.e., the identity operator) and let R_0 be the L^2 adjoint of $R_0^\top$, i.e.,

$$
(v_0, R_0 w) := (R_0^\top v_0, w) \quad \text{for all} \quad w \in V^h, v_0 \in V_0.
$$

In matrix form, the two-level additive Schwarz preconditioner can now simply be written as

$$
\mathbf{M}_{AS,2}^{-1} = \sum_{i=0}^{N} \mathbf{R}_i^T \mathbf{B}_i^{-1} \mathbf{R}_i, \quad \text{where} \quad \mathbf{B}_i := \mathbf{R}_i \mathbf{B} \mathbf{R}_i^T, \quad i = 0, \ldots, N, \tag{2.23}
$$

and where $\mathbf{R}_0$ is the matrix representation of R_0 with respect to a specified basis of the coarse space V_0. The preconditioned system (2.17) then reads

$$
\mathbf{M}_{AS,2}^{-1} \mathbf{B} \mathbf{u} \ = \ \mathbf{M}_{AS,2}^{-1} \mathbf{f}. \tag{2.24}
$$

The convergence of GMRES for (2.24) was analysed in [9] for a classical piecewise linear coarse space (see also [38, Chapter 11]). Under an a priori regularity assumption on the solution w of the adjoint problem (see the discussion just after Assumption 2.3(ii)), and for sufficiently small coarse mesh size H, it was shown in [9] that GMRES applied to (2.24) converges in a number of iterations that is independent of h and depends only on H/δ, where $\delta > 0$ is the size of the overlap of the subdomains. With the classical coarse space, the convergence rate of GMRES also depends in general on the coefficients appearing in the PDE, in particular on $a_{\max}/a_{\min}$. In order to achieve robustness to strong heterogeneities in the coefficients and to avoid a priori assumptions on the regularity of the adjoint solution, we will introduce in Section 2.3 an alternative coarse space V_0.

Before leaving this subsection we introduce some notation which will be useful in the analysis later. For each $i = 0, \ldots, N$ we define projections $T_i \colon V^h \to V_i$, by

$$b_{\Omega_i}(T_i u, v) = b(u, R_i^\top v), \quad \text{for all} \quad v \in V_i, \tag{2.25}$$

where $\Omega_0 := \Omega$. Conditions which ensure that the operators T_i are well-defined are presented in Section 3.2. Given the operators T_i, we can define the operator $T \colon V^h \to V^h$ as

$$T := \sum_{i=0}^{N} R_i^\top T_i, \tag{2.26}$$

which represents the preconditioner (2.23) in the following sense.

Proposition 2.6. *For any $u, v \in V^h$ with nodal vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have*

$$\langle \mathbf{M}_{AS,2}^{-1} \mathbf{B}\mathbf{u}, \mathbf{v} \rangle_{\mathbf{A}} = (Tu, v)_a, \tag{2.27}$$

where $\langle \cdot, \cdot \rangle_{\mathbf{A}}$ denotes the inner product on $\mathbb{R}^n$ induced by the matrix $\mathbf{A}_{i,j} = a(\phi_i, \phi_j)$.

2.3 GenEO coarse space

In this subsection we introduce the GenEO coarse space first proposed in [37].

For $1 \leq j \leq N$, let $\tilde{n}_j$ be the dimension of $\widetilde{V}_j$ and let $\{\phi_1^j, \ldots, \phi_{\tilde{n}_j}^j\}$ be a nodal basis of $\widetilde{V}_j$. We first define the local partition of unity operator.

Definition 2.7 (Partition of unity). *Let $\mathrm{dof}(\Omega_j)$ denote the internal degrees of freedom on subdomain Ω_j. For any degree of freedom i, let μ_i denote the number of subdomains for which i is an internal degree of freedom, i.e.,*

$$\mu_i := \#\{j \;:\; 1 \leq j \leq N, \;\; i \in \mathrm{dof}(\Omega_j)\}. \tag{2.28}$$

Then, for $1 \leq j \leq N$, the local partition of unity operator $\Xi_j \colon \widetilde{V}_j \to V_j$ is defined by

$$\Xi_j(v) := \sum_{i \in \mathrm{dof}(\Omega_j)} \frac{1}{\mu_i} v_i \phi_i^j, \quad \text{for all} \;\; v = \sum_{i=1}^{\tilde{n}_j} v_i \phi_i^j \in \widetilde{V}_j. \tag{2.29}$$

The operators Ξ_j form a partition of unity in the following sense [37]:

$$\sum_{j=1}^{N} R_j^\top \Xi_j(v|_{\Omega_j}) = v, \quad \text{for all} \;\; v \in V^h. \tag{2.30}$$

The local generalized eigenproblems that form the basis of the GenEO coarse space are now introduced.

Definition 2.8. *For each $j = 1, \ldots, N$ we define the following generalized eigenvalue problem:*

$$\textit{Find} \quad (p, \lambda) \in \widetilde{V}_j \backslash \{0\} \times \mathbb{R} \quad \textit{such that} \quad a_{\Omega_j}(p, v) = \lambda\, a_{\Omega_j}(\Xi_j(p), \Xi_j(v)), \quad \text{for all} \;\; v \in \widetilde{V}_j, \tag{2.31}$$

where Ξ_j is the local partition of unity operator from Definition 2.7.

Note that there is some flexibility in the choice of the bilinear form on the right hand side of (2.31). The one chosen here is not the same as the one in the original paper [37]. However, it is identical to the one considered in [12, Chapter 7]. See also [3] for an analysis of both choices as well as a further one.

Definition 2.9. *(GenEO coarse space). For each $j = 1, \dots, N$, let $(p_l^j)_{l=1}^{m_j}$ be the eigenfunctions corresponding to the m_j smallest eigenvalues in (2.31). We then define the coarse space V_0 by*

$$V_0 := \operatorname{span}\{R_j^\top \Xi_j(p_l^j) \colon l = 1, \dots, m_j \text{ and } j = 1, \dots, N\}. \tag{2.32}$$

In [37], a two-level overlapping Schwarz preconditioner with this GenEO coarse space was analysed for self-adjoint and coercive problems, in particular for (2.11) with $\mathbf{b} = 0$ and $c(x) \geq 0$. It was shown there that a conjugate gradient (CG) iteration applied to the preconditioned system (2.26) converges in a number of iterations that is independent of h, H, and δ, as well as of the variation of the coefficients $A(x)$ and $c(x)$, in particular independent of $a_{\max}/a_{\min}$.

The result in the self-adjoint and coercive case holds without any constraints on H or Θ. This cannot be expected in the indefinite or non-self-adjoint case. However, by combining the ideas from [9] and [37], a coefficient-robust theory for two-level overlapping Schwarz with GenEO coarse space can still be developed, as we will see in the following two sections.

3 Theoretical tools

3.1 Basic properties of GenEO coarse space

In this subsection, we give some important properties of the GenEO coarse space for subsequent use. The following lemma gives an error estimate for a local projection operator approximating a function $v \in \widetilde{V}_j$ in the space spanned by the eigenfunctions in (2.31).

Lemma 3.1 ([37, Lemma 2.11])**.** *Let $j \in \{1, \dots, N\}$ and $\{(p_l^j, \lambda_l^j)\}_{l=1}^{\dim(V_j)}$ be the eigenpairs of the generalized eigenproblem (2.31). Suppose that $m_j \in \{1, \dots, \dim(V_j) - 1\}$ is such that $0 < \lambda_{m_j+1}^j < \infty$. Then the local projection operator $\Pi_{m_j}^j$, defined by*

$$\Pi_{m_j}^j v := \sum_{l=1}^{m_j} a_{\Omega_j}(\Xi_j(v), \Xi_j(p_l^j))\, p_l^j, \tag{3.1}$$

satisfies the estimates

$$\|v - \Pi_{m_j}^j v\|_{a,\Omega_j} \ \leq \ \|v\|_{a,\Omega_j}, \quad \text{and} \quad \left\|\Xi_j(v - \Pi_{m_j}^j v)\right\|_{a,\Omega_j}^2 \leq \frac{1}{\lambda_{m_j+1}^j} \left\|v - \Pi_{m_j}^j v\right\|_{a,\Omega_j}^2, \tag{3.2}$$

for all $v \in \widetilde{V}_j$.

Building on these local error estimates, we can now establish a global approximation property for the GenEO coarse space. This property is a crucial element of the robustness proof for the two-level Schwarz method for the indefinite problem. Before proceeding, we formally define the integer k_0, which measures the number of subdomains any element can belong to, as

$$k_0 = \max_{\tau \in \mathcal{T}_h} \left(\#\{\Omega_j : 1 \leq j \leq N,\ \tau \subset \Omega_j\} \right). \tag{3.3}$$

Lemma 3.2. *Under the same conditions as Lemma 3.1, let $v \in V^h$. Then*

$$\inf_{z \in V_0} \|v - z\|_a^2 \leq k_0^2 \Theta \|v\|_a^2, \quad \text{where} \quad \Theta = \left(\min_{1 \leq j \leq N} \lambda_{m_j+1}^j \right)^{-1}. \tag{3.4}$$

Proof. Given $v \in V^h$, we define

$$z_0 := \sum_{j=1}^{N} \Xi_j(\Pi^j_{m_j} v|_{\Omega_j}). \tag{3.5}$$

Then, making use of (2.30), we have

$$\|v - z_0\|_a^2 = \left\| \sum_{j=1}^{N} \Xi_j(v|_{\Omega_j} - \Pi^j_{m_j} v|_{\Omega_j}) \right\|_a^2 \leq k_0 \sum_{j=1}^{N} \|\Xi_j(v|_{\Omega_j} - \Pi^j_{m_j} v\Omega_j)\|^2_{a,\Omega_j}, \tag{3.6}$$

where in the last step we used a standard estimate, e.g., [37, Equation (2.11)]. Then, using Lemma 3.1, we have

$$\begin{aligned} \|v - z_0\|_a^2 &\leq k_0 \sum_{j=1}^{N} \|\Xi_j(v - \Pi^j_{m_j} v)\|^2_{a,\Omega_j} \leq k_0 \sum_{j=1}^{N} \frac{1}{\lambda^j_{m_j+1}} \left\|v - \Pi^j_{m_j} v\right\|^2_{a,\Omega_j} \\ &\leq k_0 \sum_{j=1}^{N} \frac{1}{\lambda^j_{m_j+1}} \|v\|^2_{a,\Omega_j} \leq k_0^2 \max_{1\leq j\leq N} \frac{1}{\lambda^j_{m_j+1}} \|v\|^2_a, \end{aligned} \tag{3.7}$$

which implies (3.4) as desired. □

The following lemma can be proved by combining Lemma 2.9 of [37] with Lemma 3.1 above. It shows that the GenEO coarse space combined with the local finite element subspaces admit a stable decomposition. Such a property is the usual tool for bounding the condition number of the two-level Schwarz preconditioner in the classical positive definite case. This property is also a key ingredient in this work, since the indefinite problem is treated as a perturbation of the associated positive definite problem in the convergence analysis.

Lemma 3.3 (Stable decomposition)**.** *Let $v \in V^h$, then the decomposition*

$$z_0 := \sum_{j=1}^{N} \Xi_j(\Pi^j_{m_j} v|_{\Omega_j}), \quad z_j := \Xi_j(v|_{\Omega_j} - \Pi^j_{m_j} v|_{\Omega_j}), \quad \text{for } j = 1, \dots, N, \tag{3.8}$$

satisfies $v = \sum_{j=0}^{N} z_j$ *and*

$$\|z_0\|_a^2 + \sum_{j=1}^{N} \|z_j\|^2_{a,\Omega_j} \leq \beta_0^2(k_0\Theta^{1/2})\|v\|_a^2, \quad \textit{where} \quad \beta_0(z) = 2\sqrt{1+z^2}, \quad \textit{for } z \in \mathbb{R}^+. \tag{3.9}$$

Remark. The result from [37] actually yields (3.9) with $\beta_0^2(k_0\Theta^{1/2})$ replaced by $2+k_0(2k_0+1)\Theta$. We give the estimate in the form (3.9) for simplicity, showing that the key parameter is $k_0\Theta^{1/2}$.

As is well-known, the stable decomposition of Lemma 3.3 provides the theoretical basis for the analysis of preconditioners for systems arising from (1.1) in the special case when $\mathbf{b}$ and c vanish. This can be expressed by introducing the projection operators $P_i \colon V^h \to V^i$ such that

$$a_{\Omega_i}(P_i u, v) = a(u, R_i^\top v), \quad \forall v \in V^i, \quad i = 0, 1, \dots, N, \tag{3.10}$$

These operators are well-defined, as shown in [38, Section 2.2]. Moreover, defining $P \colon V^h \to V^h$ as

$$P := \sum_{i=0}^{N} R_i^\top P_i, \tag{3.11}$$

then using the stable decomposition property (3.9), it was proved in [38, Section 2.3] that any $u \in V^h$ satisfies

$$\beta_0^{-2}(k_0\Theta^{1/2})a(u,u) \le a(Pu,u), \tag{3.12}$$

and

$$\sum_{i=0}^{N} \|P_i u\|_{a,\Omega_i}^2 \le (k_0+1)\|u\|_a^2. \tag{3.13}$$

3.2 Solvability of subproblems

In this subsection, we prove that the operators T_i are well-defined by (2.25). Stability properties of these operators are then analysed in Section 3.3. The analysis builds again on the result of [34], given here in Lemma 2.4. Since the underlying PDE is not coercive, this requires some discussion and the imposition of some resolution conditions. Recalling items (ii) and (iv) from Assumption 2.1, we can define

$$c_{\text{inf}}^- = \operatorname{essinf}\{c^-(x) : \; x \in \Omega\}.$$

Note that, if $c_{\text{inf}}^- \ge 0$, then $c = c^- + c^+$ is almost everywhere non-negative.

For the well-posedness and stability of the local problems on each subdomain Ω_i, we will need a coefficient-dependent resolution condition on H. We specify this via the function C_0, depending on $\mathbf{b}$ and c^-, given by

$$C_0(\mathbf{b}, c^-) := \left(\frac{1}{2}\|\mathbf{b}\|_\infty^2 \; + \; \max\{-c_{\text{inf}}^-, 0\}\right)^{1/2}, \tag{3.14}$$

and note that C_0 vanishes when $c_{\text{inf}}^- = 0$ and $\mathbf{b} = \mathbf{0}$.

At various points in the following analysis we will use the standard inequality

$$pq \;\le\; \frac{1}{2\varepsilon}p^2 + \frac{\varepsilon}{2}q^2, \quad \text{valid for all} \quad p, q \in \mathbb{R} \quad \text{and} \quad \varepsilon > 0. \tag{3.15}$$

Lemma 3.4 (T_i is well-defined for each $i = 1, \ldots, N$)**.** *Suppose $H\,C_0(\mathbf{b}, c^-) < 1/2$. Then the operators T_i, $i = 1, \ldots, N$, are well-defined. In particular, each T_i is well-defined without a condition on H when $c_{\text{inf}}^- \ge 0$ and $\mathbf{b} = \mathbf{0}$.*

Proof. We need to show that, for each $u \in V^h$ and $i = 1, \ldots, N$, the discrete problem

$$\text{Find} \quad \phi_i \in V_i \quad \text{such that} \quad b_{\Omega_i}(\phi_i, v) = b(u, R_i^\top v), \quad \forall v \in V_i \tag{3.16}$$

has a unique solution ϕ_i. Since (3.16) is equivalent to a square linear system, it is sufficient to prove that, when $u = 0$, we only have the trivial solution. To do so, let $\phi_i \in V_i$ solve the problem

$$b_{\Omega_i}(\phi_i, v) = 0, \quad \text{for all} \quad v \in V_i. \tag{3.17}$$

Then, taking $v = \phi_i$ in (3.17) and using the fact that $-c^-(x) \le \max\{-c_{\text{inf}}^-, 0\}$ for almost all $x \in \Omega$ and then (2.9), we find

$$\begin{aligned}
\|\phi_i\|_{a,\Omega_i}^2 &= -(\mathbf{b}\cdot\nabla\phi_i, \phi_i)_{\Omega_i} - (c^-\phi_i, \phi_i)_{\Omega_i} \\
&\le \|\mathbf{b}\|_{L^\infty(\Omega_i)}\|\nabla\phi_i\|_{L^2(\Omega_i)}\|\phi_i\|_{L^2(\Omega_i)} + \max\{-c_{\text{inf}}^-, 0\}\|\phi_i\|_{L^2(\Omega_i)}^2 \\
&\le \frac{H}{\sqrt{2}}\|\mathbf{b}\|_\infty\|\nabla\phi_i\|_{L^2(\Omega_i)}^2 + \frac{H^2}{2}\max\{-c_{\text{inf}}^-, 0\}\|\nabla\phi_i\|_{L^2(\Omega_i)}^2 \\
&\le \left(\frac{H}{\sqrt{2}}\|\mathbf{b}\|_\infty + \frac{H^2}{2}\max\{-c_{\text{inf}}^-, 0\}\right)\|\phi_i\|_{a,\Omega_i}^2.
\end{aligned} \tag{3.18}$$

Now it is easy to see that the condition $H\, C_0(\mathbf{b}, c^-) < 1/2$ ensures that

$$\frac{H}{\sqrt{2}}\|\mathbf{b}\|_\infty < \frac{1}{2} \quad \text{and} \quad \frac{H^2}{2}\max\{-c^-_{\mathrm{inf}}, 0\} < 1/8.$$

Thus (3.18) implies

$$\|\phi_i\|^2_{a,\Omega_i} \le \frac{5}{8}\|\phi_i\|^2_{a,\Omega_i},$$

and it follows that $\phi_i = 0$. $\square$

Now, to show that T_0 is also well-defined, we need to impose a resolution condition on Θ, specified using the function $C_0(\mathbf{b}, c^-)$, together with another coefficient-dependent function

$$C_1(\mathbf{b}, c^-) := 1 + \|\mathbf{b}\|_\infty + \|c^-\|_\infty. \tag{3.19}$$

Recall also the quantities k_0, defined in (3.3), and C^*_{stab}, defined in Assumption 2.3(ii).

Lemma 3.5 (T_0 is well-defined)**.** *Suppose that*

$$[(1 + C^*_{\mathrm{stab}})\, C_0(\mathbf{b}, c^-)\, C_1(\mathbf{b}, c^-) k_0]\Theta^{1/2} < 1/\sqrt{2}.$$

Then there exists $h_* > 0$ such that, for all $h \le h_*$, the operator T_0 is well-defined. In particular, T_0 is well-defined, without condition on Θ, when $c^-_{\mathrm{inf}} \ge 0$ and $\mathbf{b} = \mathbf{0}$.

Proof. Analogously to Lemma 3.4, it is sufficient to consider any solution ϕ to the problem

$$\text{Find} \quad \phi_0 \in V_0 \quad \text{such that} \quad b(\phi_0, v) = 0, \quad \text{for all} \quad v \in V_0, \tag{3.20}$$

and then to show that $\phi_0 = 0$. To do so, given ϕ_0, we first note that

$$\begin{aligned} 0 = b(\phi_0, \phi_0) \;&=\; a(\phi_0, \phi_0) + (c^-\phi_0, \phi_0) + (\mathbf{b}\cdot\nabla\phi_0, \phi_0) \\ &\ge \|\phi_0\|^2_a + c^-_{\mathrm{inf}}\|\phi_0\|^2 + (\mathbf{b}\cdot\nabla\phi_0, \phi_0) \\ &\ge \|\phi_0\|^2_a + c^-_{\mathrm{inf}}\|\phi_0\|^2 - \|\mathbf{b}\|_\infty\|\nabla\phi_0\|\,\|\phi_0\| \\ &\ge \|\phi_0\|^2_a + c^-_{\mathrm{inf}}\|\phi_0\|^2 - \frac{1}{2}\left(\|\phi_0\|^2_a + \|\mathbf{b}\|^2_\infty\|\phi_0\|^2\right) \\ &= \frac{1}{2}\|\phi_0\|^2_a - \left(\frac{1}{2}\|\mathbf{b}\|^2_\infty - c^-_{\mathrm{inf}}\right)\|\phi_0\|^2 \\ &\ge \frac{1}{2}\|\phi_0\|^2_a - C_0(\mathbf{b}, c^-)^2\|\phi_0\|^2, \end{aligned} \tag{3.21}$$

where in the third step we used (3.15) and (2.9), and in the final step we used the definition of $C_0(\mathbf{b}, c^-)$ in (3.14).

To finish the proof we use a duality argument to estimate $\|\phi_0\|$ from above, in terms of $\|\phi_0\|_a$. As such, let w solve the auxiliary problem:

$$\text{Find} \quad w \in H^1_0(\Omega) \quad \text{such that} \quad b(v, w) = (\phi_0, v), \quad \text{for all } v \in H^1_0(\Omega). \tag{3.22}$$

Further, let $w_h \in V^h$ denote its finite element approximation, i.e.,

$$b(v, w_h) = (\phi_0, v), \quad \text{for all} \quad v \in V^h. \tag{3.23}$$

Applying Lemma 2.4 to the adjoint problem (3.22), we see that there exists an $h_* > 0$ such that, for all $0 < h < h_*$,

$$\|w - w_h\|_a \le \|\phi_0\|\,. \tag{3.24}$$

(See also the discussion following Lemma 2.4.) Now, combining (3.24) and Assumption 2.3(ii), it follows that

$$\|w_h\|_a \leq (C^*_{\text{stab}} + 1)\|\phi_0\|. \tag{3.25}$$

Since $V_0 \subset V^h$, we can substitute $v = \phi_0$ in (3.23) and then use (3.20) and (2.9) to obtain, for any $z \in V_0$,

$$\begin{aligned}\|\phi_0\|^2 &= b(\phi_0, w_h) = b(\phi_0, w_h - z)\\ &\leq \|\phi_0\|_a \|w_h - z\|_a + \|\mathbf{b}\|_\infty \|\nabla\phi_0\| \|w_h - z\| + \|c^-\|_\infty \|\phi_0\| \|w_h - z\|\\ &\leq \left(1 + \|\mathbf{b}\|_\infty D_\Omega + \|c^-\|_\infty D_\Omega^2\right) \|\phi_0\|_a \|w_h - z\|_a.\end{aligned} \tag{3.26}$$

Thus, using Assumption 2.1(iii), we have

$$\|\phi_0\|^2 \leq C_1(\mathbf{b}, c^-)\|\phi_0\|_a \|w_h - z\|_a, \quad \text{for all} \quad z \in V_0. \tag{3.27}$$

Now, from (3.27) using Lemma 3.2 and then (3.25), we further have

$$\begin{aligned}\|\phi_0\|^2 &\leq C_1(\mathbf{b}, c^-)\|\phi_0\|_a \inf_{z\in V_0} \|w_h - z\|_a \leq C_1(\mathbf{b}, c^-)k_0\Theta^{\frac{1}{2}}\|\phi_0\|_a\|w_h\|_a\\ &\leq (1 + C^*_{\text{stab}})C_1(\mathbf{b}, c^-)k_0\Theta^{1/2}\|\phi_0\|_a\|\phi_0\|,\end{aligned} \tag{3.28}$$

and thus

$$\|\phi_0\| \leq (1 + C^*_{\text{stab}})C_1(\mathbf{b}, c^-)k_0\Theta^{1/2}\|\phi_0\|_a. \tag{3.29}$$

Finally, inserting (3.29) into (3.21), we obtain

$$0 \geq \left(1/2 - K^2\right) \|\phi_0\|_a^2,$$

where

$$K := (1 + C^*_{\text{stab}})\, C_0(\mathbf{b}, c^-)\, C_1(\mathbf{b}, c^-)k_0\Theta^{1/2}, \tag{3.30}$$

and $\phi_0 = 0$ follows necessarily when $K < 1/\sqrt{2}$. □

3.3 Stability estimates for T_i

In this subsection, we prove the stability estimates for the operators T_i that will be required in the following section to establish the robustness of the two-level additive Schwarz method.

Lemma 3.6 (Stability for T_i, $i = 1, \dots, N$)**.** *Suppose $H\, C_0(\mathbf{b}, c^-) < 1/4$. Then for all $u \in V^h$ and all $i \in \{1, \dots, N\}$*

$$\|T_i u\|_{a,\Omega_i} \leq 2\sqrt{2}\|u\|_{a,\Omega_i} + 2\sqrt{\frac{2}{3}}H\|c^-\|_{L^\infty(\Omega_i)}\, \|u\|_{L^2(\Omega_i)}. \tag{3.31}$$

Proof. Let $i \in \{1, \dots, N\}$. Using the fact that $b_{\Omega_i}(T_i u, T_i u) = b_{\Omega_i}(u, T_i u)$, we find

$$\begin{aligned}\|T_i u\|^2_{a,\Omega_i} &= a(T_i u, T_i u)\\ &= b_{\Omega_i}(T_i u, T_i u) - \left[(\mathbf{b}\cdot\nabla T_i u, T_i u) + (c^- T_i u, T_i u)_{\Omega_i}\right]\\ &= b_{\Omega_i}(u, T_i u) - \left[(\mathbf{b}\cdot\nabla T_i u, T_i u) + (c^- T_i u, T_i u)_{\Omega_i}\right]\\ &= a_{\Omega_i}(u, T_i u) + (\mathbf{b}\cdot\nabla u, T_i u)_{\Omega_i} + (c^- u, T_i u)_{\Omega_i}\\ &\qquad - \left[(\mathbf{b}\cdot\nabla T_i u, T_i u)_{\Omega_i} + (c^- T_i u, T_i u)_{\Omega_i}\right].\end{aligned} \tag{3.32}$$

We estimate each of the terms in (3.32) as follows, using (2.9) and (3.15), to obtain

$$
\begin{aligned}
a_{\Omega_i}(u, T_i u) &\leq \frac{1}{4}\|T_i u\|_{a,\Omega_i}^2 + \|u\|_{a,\Omega_i}^2, && \text{(3.33a)}\\
(\mathbf{b}\cdot\nabla u, T_i u) &\leq \|\mathbf{b}\|_{L^\infty(\Omega_i)}\|u\|_{a,\Omega_i}\|T_i u\|_{L^2(\Omega_i)} \leq \frac{H}{\sqrt{2}}\|\mathbf{b}\|_\infty\|T_i u\|_{a,\Omega_i}\|u\|_{a,\Omega_i}\\
&\leq \frac{H^2}{4}\|\mathbf{b}\|_{L^\infty(\Omega_i)}^2\|T_i u\|_{a,\Omega_i}^2 + \frac{1}{2}\|u\|_{a,\Omega_i}^2, && \text{(3.33b)}\\
\text{and}\quad (c^- u, T_i u)_{\Omega_i} &\leq \frac{H}{\sqrt{2}}\|T_i u\|_{a,\Omega_i}\,\|c^-\|_{L^\infty(\Omega_i)}\|u\|_{L^2(\Omega_i)}\\
&\leq \frac{1}{4}\|T_i u\|_{a,\Omega_i}^2 + \frac{H^2}{2}\|c^-\|_{L^\infty(\Omega_i)}^2\|u\|_{L^2(\Omega_i)}^2. && \text{(3.33c)}
\end{aligned}
$$

Combining (3.33) with (3.32) and estimating the final term in (3.32) using (3.18), we obtain

$$
\begin{aligned}
\|T_i u\|_{a,\Omega_i}^2 \leq{}& \left(\frac{1}{2} + \frac{H^2}{4}\|\mathbf{b}\|_{L^\infty(\Omega_i)}^2\right)\|T_i u\|_{a,\Omega_i}^2\\
&+ \left(\frac{3}{2}\|u\|_{a,\Omega_i}^2 + \frac{H^2}{2}\|c^-\|_{L^\infty(\Omega_i)}^2\|u\|_{L^2(\Omega_i)}^2\right)\\
&+ \left(\frac{H}{\sqrt{2}}\|\mathbf{b}\|_{L^\infty(\Omega_i)} + \frac{H^2}{2}\max\{-c_{\inf}^-, 0\}\right)\|T_i u\|_{a,\Omega_i}^2.
\end{aligned}
$$

Hence, we deduce that

$$
\begin{aligned}
\frac{1}{2}\|T_i u\|_{a,\Omega_i}^2 \leq{}& \left[\frac{H}{\sqrt{2}}\|\mathbf{b}\|_\infty + \frac{H^2}{4}\|\mathbf{b}\|_\infty^2 + \frac{H^2}{2}\max\{-c_{\inf}^-, 0\}\right]\|T_i u\|_{a,\Omega_i}^2\\
&+ \left(\frac{3}{2}\|u\|_{a,\Omega_i}^2 + \frac{H^2}{2}\|c^-\|_{L^\infty(\Omega_i)}^2\|u\|_{L^2(\Omega_i)}^2\right).
\end{aligned} \tag{3.34}
$$

Arguing as in the proof of Lemma 3.4, it is now easy to see that the assumption $H\,C_0(\mathbf{b}, c^-) < 1/4$ ensures that the term in the square brackets in (3.34) is bounded above by $5/16$. Thus we have

$$
\frac{3}{16}\|T_i u\|_{a,\Omega_i}^2 \leq \left(\frac{3}{2}\|u\|_{a,\Omega_i}^2 + \frac{H^2}{2}\|c^-\|_{L^\infty(\Omega_i)}^2\|u\|_{L^2(\Omega_i)}^2\right),
$$

from which the result (3.31) follows. □

We now examine the stability of T_0 in the following lemma, for which we need another constant, namely

$$
C_2(\mathbf{b}, c^-) := 1 + \|\mathbf{b}\|_\infty + \|\nabla\cdot\mathbf{b}\|_\infty + \|c^-\|_\infty. \tag{3.35}
$$

Lemma 3.7 (Stability for T_0). *Suppose that*

$$
(1 + C_{\text{stab}}^*)\,C_0(\mathbf{b}, c^-)\,C_1(\mathbf{b}, c^-)\,k_0\Theta^{1/2} < \frac{1}{\sqrt{2}}. \tag{3.36}
$$

Then there exists $h_* > 0$ such that, for $h \leq h_*$,

$$
\|T_0 u - u\| \leq (1 + C_{\text{stab}}^*)\,C_1(\mathbf{b}, c^-)\,k_0\Theta^{1/2}\,\|T_0 u - u\|_a, \quad \textit{for all} \quad u \in V^h. \tag{3.37}
$$

Suppose, in addition, that

$$
(1 + C_{\text{stab}}^*)\,C_1(\mathbf{b}, c^-)\,C_2(\mathbf{b}, c^-)\,k_0\Theta^{1/2} < \frac{1}{2}. \tag{3.38}
$$

Then

$$
\|u - T_0 u\|_a \leq \sqrt{2}\|u\|_a, \quad \textit{for all} \quad u \in V^h. \tag{3.39}
$$

Proof. Under condition (3.36), Lemma 3.5 ensures the existence of h_* such that, when $h \le h_*$, $T_0 \colon V^h \to V_0$ is well-defined. We then consider the auxiliary problem:

$$\text{Find} \quad w_h \in V^h \quad \text{such that} \quad b(v, w_h) = (T_0 u - u, v), \quad \text{for all} \quad v \in V^h. \tag{3.40}$$

Then, since $b(T_0 u - u, z) = 0$ for all $z \in V_0$, (3.37) follows by the same argument as used in the proof of (3.29).

By the definition of P_0, we have $a(T_0 u, u - P_0 u) = 0$. Also, since $P_0 u - T_0 u \in V^0$, we have $b(u - T_0 u, P_0 u - T_0 u) = 0$. Thus, using these relations and (2.7) (twice), we can write

$$\begin{aligned}
\|u - T_0 u\|_a^2 &= b(u - T_0 u, u - P_0 u) + \big(\mathbf{b} \cdot \nabla(u - T_0 u),\, u - T_0 u\big) - (\tilde{c}(u - T_0 u), u - T_0 u) \\
&= a(u - T_0 u, u - P_0 u) + \big(\mathbf{b} \cdot \nabla(P_0 u - T_0 u),\, u - T_0 u\big) + (\tilde{c}(u - T_0 u), T_0 u - P_0 u) \\
&= a(u, u - P_0 u) + \big(\mathbf{b} \cdot \nabla(P_0 u - T_0 u),\, u - T_0 u\big) + (\tilde{c}(u - T_0 u), T_0 u - P_0 u).
\end{aligned}$$

Hence, using (2.9), it follows that

$$\begin{aligned}
\|u - T_0 u\|_a^2 \le \|u\|_a \|u - P_0 u\|_a + \|\mathbf{b}\|_\infty \|u - T_0 u\| \|P_0 u - T_0 u\|_a \\
+ \|\tilde{c}\|_\infty \|u - T_0 u\| \|P_0 u - T_0 u\|.
\end{aligned} \tag{3.41}$$

Now, by the definition of the projection P_0, we have

$$\|u - P_0 u\|_a \le \|u\|_a, \quad \text{and} \quad \|P_0 u - T_0 u\| \le \|P_0 u - T_0 u\|_a = \|P_0(u - T_0 u)\|_a \le \|u - T_0 u\|_a. \tag{3.42}$$

Hence, using (3.37) and (3.42) in (3.41), we obtain

$$\begin{aligned}
\|u - T_0 u\|_a^2 &\le \|u\|_a^2 + (\|\mathbf{b}\|_\infty + \|\tilde{c}\|_\infty)\, \|u - T_0 u\|_a \, \|u - T_0 u\| \\
&\le \|u\|_a^2 + (1 + C^*_{\mathrm{stab}})\, C_1(\mathbf{b}, c^-)\, C_2(\mathbf{b}, c^-)\, k_0 \Theta^{1/2}\, \|u - T_0 u\|_a^2,
\end{aligned}$$

and the result follows. $\square$

4 Main results

In this section we now state and prove the main theoretical result of this paper, as well as a corollary, on the robust GMRES convergence on the preconditioned system when using the GenEO coarse space, under certain conditions on the size of $\mathbf{b}$ and c. We first recall some important quantities defined in the preceding section.

$$\begin{aligned}
k_0 &= \max_{\tau \in \mathcal{T}_h} \left(\#\{\Omega_j : 1 \le j \le N,\ \tau \subset \Omega_j\} \right), \\
\Theta &= \left(\min_{1 \le j \le N} \lambda^j_{m_j+1} \right)^{-1}, \\
C_0(\mathbf{b}, c^-) &= \left(\frac{1}{2} \|\mathbf{b}\|_\infty^2 + \max\{-c^-_{\mathrm{inf}}, 0\} \right)^{1/2}, \\
C_1(\mathbf{b}, c^-) &= 1 + \|\mathbf{b}\|_\infty + \|c^-\|_\infty, \\
C_2(\mathbf{b}, c^-) &= 1 + \|\mathbf{b}\|_\infty + \|\nabla \cdot \mathbf{b}\|_\infty + \|c^-\|_\infty, \\
\beta_0(z) &= 2\sqrt{1 + z^2}.
\end{aligned} \tag{4.1}$$

Theorem 4.1. *Assume that $h \le h_*$, where h_* is as defined in Lemma 3.5. Suppose H and Θ are chosen so that*

$$s = s(\Theta) := 2\sqrt{2}(1 + C^*_{\mathrm{stab}})\, C_1(\mathbf{b}, c^-)\, C_2(\mathbf{b}, c^-)\, k_0^{3/2}\, \beta_0^2(k_0 \Theta^{1/2})\, \Theta^{1/2} < 1, \tag{4.2}$$

and

$$t = t(H, \Theta) := 16 H\, C_1(\mathbf{b}, c^-)\, k_0\, \beta_0^2(k_0 \Theta^{1/2}) < 1. \tag{4.3}$$

Then, for all $u \in V_h$,

$$c_1(H,\Theta)\, a(u,u) \le a(Tu,u), \tag{4.4}$$

and

$$a(Tu,Tu) \le c_2(H,\Theta)\, a(u,u), \tag{4.5}$$

where $c_1(H,\Theta)$ *and* $c_2(H,\Theta)$ *are given by*

$$\begin{aligned} c_1(H,\Theta) &= \beta_0^{-2}(k_0\Theta^{1/2})(1-\max\{t,\, s\}),\\ c_2(H,\Theta) &= 12+32k_0^2. \end{aligned} \tag{4.6}$$

Proof. We begin by showing that the assumptions (4.2) and (4.3) ensure that the hypotheses of Lemmas 3.6 and 3.7 hold. First, since $k_0 \ge 1$ and $\beta_0^2(k_0\Theta^{1/2}) \ge 2$, (4.2) implies

$$(1+C^*_{\mathrm{stab}})\, C_1(\mathbf{b},c^-)\, C_2(\mathbf{b},c^-)\, k_0\Theta^{1/2} < 1/(4\sqrt{2}). \tag{4.7}$$

Further, by definition (3.14) and using (3.15) we have

$$\begin{aligned} C_0^2(\mathbf{b},c^-) &\le \frac{1}{2}\|\mathbf{b}\|_\infty^2 + \|c^-\|_\infty \le \frac{1}{2}\left(\|\mathbf{b}\|_\infty^2 + \|c^-\|_\infty^2 + 1\right)\\ &\le \frac{1}{2}\left(\|\mathbf{b}\|_\infty + \|c^-\|_\infty + 1\right)^2 \le \frac{1}{2}C_2^2(\mathbf{b},c^-), \end{aligned}$$

and so (4.7) implies $(1+C^*_{\mathrm{stab}})\, C_1(\mathbf{b},c^-)\, C_0(\mathbf{b},c^-)\, k_0\Theta^{1/2} < 1/8$, and both hypotheses (3.36) and (3.38) of Lemma 3.7 are satisfied.

Likewise, hypothesis (4.3) implies $H\, C_1(\mathbf{b},c^-) < 1/32$ and so we have

$$\max\{H\|\mathbf{b}\|_\infty, H\|c^-\|_\infty\} < 1/32. \tag{4.8}$$

Combining (4.8) with a straightforward calculation and using $H \le \mathrm{diam}(\Omega) \le 1$, shows that the hypothesis of Lemma 3.6 holds. In the remainder of the proof we use the results of Lemmas 3.6 and 3.7 without further justification.

Let $u \in V_h$, we first prove (4.4). Using (3.11),(3.12), (2.6) and the definition of the projections T_i, we obtain

$$\begin{aligned} \beta_0^{-2}(k_0\Theta^{1/2})a(u,u) &\le \sum_{i=0}^N a(u, R_i^\top P_i u)\\ &= \sum_{i=0}^N b(u, R_i^\top P_i u) - \sum_{i=0}^N \left[\left(\mathbf{b}\cdot\nabla u, R_i^\top P_i u\right) + \left(c^- u, R_i^\top P_i u\right)\right]\\ &= \sum_{i=0}^N b(R_i^\top T_i u, R_i^\top P_i u) - \sum_{i=0}^N \left[\left(\mathbf{b}\cdot\nabla u, R_i^\top P_i u\right) + \left(c^- u, R_i^\top P_i u\right)\right]\\ &= \sum_{i=0}^N a(R_i^\top T_i u, R_i^\top P_i u) + \sum_{i=0}^N \left[\left(\mathbf{b}\cdot\nabla(R_i^\top T_i u - u),\, R_i^\top P_i u\right) + \left(c^-(R_i^\top T_i u - u), R_i^\top P_i u\right)\right]. \end{aligned}$$

Now, since

$$a(R_i^\top T_i u, R_i^\top P_i u) = a_{\Omega_i}(T_i u, P_i u) = a_{\Omega_i}(T_i u, u) = a(R_i^\top T_i u, u),$$

we have, by definition (2.26) of T,

$$\begin{aligned} a(u,u) &\le \beta_0^2(k_0\Theta^{1/2})\, a(Tu,u)\\ &\quad + \beta_0^2(k_0\Theta^{1/2}) \sum_{i=0}^N \left[\left(\mathbf{b}\cdot\nabla(R_i^\top T_i u - u),\, R_i^\top P_i u\right) + \left(c^-(R_i^\top T_i u - u), R_i^\top P_i u\right)\right]. \end{aligned} \tag{4.9}$$

We proceed by bounding the sum in (4.9) in terms of $a(u,u)$.

First, we consider the summand in (4.9) corresponding to $i=0$. Integrating by parts as in (2.7), recalling that $R_0^\top$ is the identity operator, and using (2.16), we obtain

$$\begin{aligned}
&\big(\mathbf{b}\cdot\nabla(R_0^\top T_0u-u),R_0^\top P_0u\big)+\big(c^-(R_0^\top T_0u-u),R_0^\top P_0u\big)\\
&\quad=-\big(\mathbf{b}\cdot\nabla(R_0^\top P_0u),R_0^\top T_0u-u\big)+\big(\tilde c(R_0^\top T_0u-u),R_0^\top P_0u\big)\\
&\quad\le\|\mathbf{b}\|_\infty\|u-T_0u\|\|P_0u\|_a+\|\tilde c\|_\infty\|u-T_0u\|\|P_0u\|\\
&\quad\le C_2(\mathbf{b},c^-)\|u-T_0u\|\|P_0u\|_a\\
&\quad\le(1+C^*_{\mathrm{stab}})\,C_2(\mathbf{b},c^-)\,C_1(\mathbf{b},c^-)\,k_0\Theta^{1/2}\|u-T_0u\|_a\|P_0u\|_a\\
&\quad\le\sqrt2(1+C^*_{\mathrm{stab}})\,C_2(\mathbf{b},c^-)\,C_1(\mathbf{b},c^-)\,k_0\Theta^{1/2}\|u\|_a\|P_0u\|_a,
\end{aligned}$$

where the last three estimates are obtained using (3.35), (3.37), and (3.39). Using this, together with the definition of s in assumption (4.2) and recalling $k_0\ge1$, we obtain

$$\begin{aligned}
&\beta_0^2(k_0\Theta^{1/2})\big[\big(\mathbf{b}\cdot\nabla(R_0^\top T_0u-u),R_0^\top P_0u\big)+\big(c^-(R_0^\top T_0u-u),R_0^\top P_0u\big)\big]\\
&\quad\le\sqrt2\,C_2(\mathbf{b},c^-)\,C_1(\mathbf{b},c^-)\,\beta_0^2(k_0\Theta^{1/2})\,(1+C^*_{\mathrm{stab}})\,k_0\Theta^{1/2}\|u\|_a\|P_0u\|_a\\
&\quad=\frac{s}{2k_0^{1/2}}\|u\|_a\|P_0u\|_a\le\frac{s}{\sqrt2(k_0+1)^{1/2}}\|u\|_a\|P_0u\|_a.
\end{aligned}\tag{4.10}$$

Next, we consider the summands in (4.9) corresponding to $i=1,\dots,N$. Using (2.16) and the fact that $R_i^\top$ corresponds to the zero extension from Ω_i, we obtain

$$\begin{aligned}
&\big(\mathbf{b}\cdot\nabla(R_i^\top T_iu-u),R_i^\top P_iu\big)+\big(c^-(R_i^\top T_iu-u),R_i^\top P_iu\big)\\
&\quad\le\big(\|\mathbf{b}\|_{L^\infty(\Omega_i)}\|u-T_iu\|_{a,\Omega_i}+\|c^-\|_{L^\infty(\Omega_i)}\|u-T_iu\|_{L^2(\Omega_i)}\big)\,\|P_iu\|_{L^2(\Omega_i)}.
\end{aligned}\tag{4.11}$$

Furthermore, by (3.31) and (2.16) again and noting that (4.8) implies $H\|c^-\|_{L^\infty(\Omega_i)}\le1$, we have

$$\|u-T_iu\|_{a,\Omega_i}\le4\|u\|_{a,\Omega_i}+2\|u\|_{L^2(\Omega_i)}\quad\text{and}\quad\|u-T_iu\|_{L^2(\Omega_i)}\le2\|u\|_{L^2(\Omega_i)}+\sqrt3\|u\|_{a,\Omega_i}.$$

Inserting these estimates into (4.11), we obtain

$$\begin{aligned}
&\big(\mathbf{b}\cdot\nabla(R_i^\top T_iu-u),R_i^\top P_iu\big)+\big(c^-(R_i^\top T_iu-u),R_i^\top P_iu\big)\\
&\quad\le\big(\|\mathbf{b}\|_{L^\infty(\Omega_i)}+\|c^-\|_{L^\infty(\Omega_i)}\big)\,\big(4\|u\|_{a,\Omega_i}+2\|u\|_{L^2(\Omega_i)}\big)\,\|P_iu\|_{L^2(\Omega_i)}.
\end{aligned}\tag{4.12}$$

Now, using Lemma 2.5 again, together with the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
&\sum_{i=1}^N\big[\big(\mathbf{b}\cdot\nabla(R_i^\top T_iu-u),R_i^\top P_iu\big)+\Big(c^-(R_i^\top T_iu-u),R_i^\top P_iu\Big)\big]\\
&\qquad\le4H\,C_1(\mathbf{b},c^-)\left(\sum_{i=1}^N\big(\|u\|_{L^2(\Omega_i)}+\|u\|_{a,\Omega_i}\big)^2\right)^{1/2}\left(\sum_{i=1}^N\|P_iu\|^2_{a,\Omega_i}\right)^{1/2}\\
&\qquad\le8H\,C_1(\mathbf{b},c^-)\left(\sum_{i=1}^N\|u\|^2_{a,\Omega_i}\right)^{1/2}\left(\sum_{i=1}^N\|P_iu\|^2_{a,\Omega_i}\right)^{1/2}\\
&\qquad\le8Hk_0^{1/2}\,C_1(\mathbf{b},c^-)\,\|u\|_a\left(\sum_{i=1}^N\|P_iu\|^2_{a,\Omega_i}\right)^{1/2}.
\end{aligned}\tag{4.13}$$

Hence, using (4.3), we obtain

$$\begin{aligned}
&\beta_0^2(k_0\Theta^{1/2})\sum_{i=1}^N\big[\big(\mathbf{b}\cdot\nabla(R_i^\top T_iu-u),\,R_i^\top P_iu\big)+\big(c^-(R_i^\top T_iu-u),R_i^\top P_iu\big)\big]\\
&\qquad\le\frac{t}{\sqrt2(k_0+1)^{1/2}}\|u\|_a\left(\sum_{i=1}^N\|P_iu\|^2_{a,\Omega_i}\right)^{1/2}.
\end{aligned}\tag{4.14}$$

It then follows from (4.10), (4.14), that

$$\begin{aligned}\beta_0^2(k_0\Theta^{1/2})&\sum_{i=0}^{N}\left[\left(\mathbf{b}\cdot\nabla(R_i^\top T_i u-u),\, R_i^\top P_i u\right)+\left(c^-(R_i^\top T_i u-u), R_i^\top P_i u\right)\right]\\
&\le \frac{\max\{s,t\}}{\sqrt{2}(k_0+1)^{1/2}}\left(\|u\|_a\|P_0 u\|_a+\|u\|_a\left(\sum_{i=1}^{N}\|P_i u\|_{a,\Omega_i}^2\right)^{\frac{1}{2}}\right)\\
&\le \frac{\max\{s,t\}}{(k_0+1)^{1/2}}\|u\|_a\left(\sum_{i=0}^{N}\|P_i u\|_{a,\Omega_i}^2\right)^{\frac{1}{2}}\le \max\{s,t\}\|u\|_a^2. \qquad (4.15)\end{aligned}$$

Now, inserting (4.15) into (4.9), we see that

$$a(u,u)\le \beta_0^2(k_0\Theta^{1/2})a(Tu,u)+\max\{s,t\}\|u\|_a^2, \qquad (4.16)$$

which implies that

$$\beta_0^{-2}(k_0\Theta^{1/2})\left(1-\max\{s,t\}\right)a(u,u)\le a(Tu,u), \qquad (4.17)$$

thus proving (4.4).

Now, to prove (4.5), we observe that

$$a(Tu,Tu)=\left\|\sum_{i=0}^{N}R_i^\top T_i u\right\|_a^2\le 2\|T_0 u\|_a^2+2\left\|\sum_{i=1}^{N}R_i^\top T_i u\right\|_a^2, \qquad (4.18)$$

Further, by (3.39), we have

$$\|T_0 u\|_a^2\le 6\|u\|_a^2, \qquad (4.19)$$

and in addition, by Lemma 3.6, we obtain

$$\begin{aligned}\left\|\sum_{i=1}^{N}R_i^\top T_i u\right\|_a^2 &\le k_0\sum_{i=1}^{N}\|T_i u\|_{a,\Omega_i}^2\le 2k_0\sum_{i=1}^{N}\left(6\|u\|_{a,\Omega_i}^2+2\|u\|_{L^2(\Omega_i)}^2\right)\\
&\le 2k_0^2\left(6\|u\|_a^2+2\|u\|_{L^2(\Omega)}^2\right)\le 16k_0^2\|u\|_a^2. \qquad (4.20)\end{aligned}$$

Combining (4.18)–(4.20), we finally determine that

$$a(Tu,Tu)\le\left(12+32k_0^2\right)\|u\|_a^2, \qquad (4.21)$$

thus completing the proof of Theorem 4.1. □

Using Theorem 4.1 and the error estimates for GMRES in [14], we obtain the following key result as a corollary.

Corollary 4.2. *Under the assumptions of Theorem 4.1, if GMRES in the $\langle\cdot,\cdot\rangle_{\mathbf{A}}$-inner product is applied to solve the preconditioned system (2.24) then, after m steps, the norm of the residual is bounded as*

$$\|\mathbf{r}^m\|_{\mathbf{A}}^2\le\left(1-\frac{c_1^2}{c_2^2}\right)^m\|\mathbf{r}^0\|_{\mathbf{A}}^2, \qquad (4.22)$$

where $\mathbf{r}^0$ is the initial residual.

Proof. The estimates (4.4) and (4.5), together with (2.27) imply

$$\left|\langle\mathbf{M}_{AS,2}^{-1}\mathbf{B}\mathbf{u},\mathbf{u}\rangle_{\mathbf{A}}\right|\ge c_1(H_0,\Lambda_0)\,\|\mathbf{u}\|_{\mathbf{A}}^2 \quad\text{and}\quad \|\mathbf{M}_{AS,2}^{-1}\mathbf{B}\|_{\mathbf{A}}^2\le c_2.$$

The result follows directly from the Elman theory [14]. □

Restricting to the special case $\mathbf{b} = \mathbf{0}$ and $c = 0$, we obtain essentially the same result as in [37], which is interesting because the latter makes use of 'self-adjoint' technology, working with eigenvalue and condition number estimates and does not estimate the field of values as in [14].

The fact that domain decomposition theory usually provides theoretical estimates in the energy inner product, while GMRES is normally applied with respect to the Euclidean inner product makes the estimate in the previous corollary slightly impractical. However this can be converted to a statement about standard GMRES using the following norm equivalence argument (see also [22, Corollary 5.8]).

Lemma 4.3. *Suppose we are solving the linear system*

$$\mathbf{C}\mathbf{x} = \mathbf{c} \tag{4.23}$$

on $\mathbb{R}^n$ *and* $\langle\cdot,\cdot\rangle_1$ *and* $\langle\cdot,\cdot\rangle_2$ *are two inner products on* $\mathbb{R}^n$ *with associated norms* $\|\cdot\|_1$ *and* $\|\cdot\|_2$*. The corresponding equivalence constants are denoted by* $\underline{\gamma}, \overline{\gamma}$*, so that*

$$\underline{\gamma}\|\mathbf{x}\|_1 \le \|\mathbf{x}\|_2 \le \overline{\gamma}\|\mathbf{x}\|_1, \quad \textit{for all} \quad \mathbf{x} \in \mathbb{R}^n. \tag{4.24}$$

Given a fixed initial guess $\mathbf{x}^0 \in \mathbb{R}^n$ *with residual* $\mathbf{r}^0$*, consider the sequences of iterates* $\mathbf{x}_1^m$ *and* $\mathbf{x}_2^m$*,* $m = 0, 1, 2, \ldots,$ *with residuals* $\mathbf{r}_1^m$ *and* $\mathbf{r}_2^m$ *that minimise the residual of* (4.23) *over the Krylov subspace spanned by* C *in the norms* $\|\cdot\|_1$ *and* $\|\cdot\|_2$ *respectively. Suppose also that the sequence* $\mathbf{r}_1^m$ *enjoys the relative residual estimate*

$$\|\mathbf{r}_1^m\|_1 \le \sigma^m \|\mathbf{r}^0\|_1, \quad \textit{for some} \quad \sigma < 1.$$

Then we have

$$\|\mathbf{r}_2^{m+\Delta m}\|_2 \le \sigma^m \|\mathbf{r}^0\|_2, \quad \textit{for all} \quad \Delta m \ge \frac{\log(\overline{\gamma}/\underline{\gamma})}{\log(\sigma^{-1})}. \tag{4.25}$$

Proof. Since $\mathbf{x}_2$ minimises the residual in the norm $\|\cdot\|_2$ and $\mathbf{r}_1^0 = \mathbf{r}_2^0 = \mathbf{c} - C\mathbf{x}^0$, it follows from (4.24) that

$$\frac{\|\mathbf{r}_2^{m+\Delta m}\|_2}{\|\mathbf{r}_2^0\|_2} \le \frac{\|\mathbf{r}_1^{m+\Delta m}\|_2}{\|\mathbf{r}_1^0\|_2} \le \frac{\overline{\gamma}}{\underline{\gamma}} \frac{\|\mathbf{r}_1^{m+\Delta m}\|_1}{\|\mathbf{r}_1^0\|_1} \le \frac{\overline{\gamma}}{\underline{\gamma}} \sigma^{m+\Delta m}.$$

Thus, we can deduce the bound in (4.25) if $\Delta m \log(\sigma^{-1}) \ge \log(\overline{\gamma}/\underline{\gamma})$. □

Corollary 4.4. *Assuming the mesh sequence* $\mathcal{T}_h$ *is quasiuniform, then with an additional number* Δm *of iterations, which grows at most proportionally to* $\log(a_{\max}) + \log(h^{-1})$*, GMRES applied in the Euclidean inner product (i.e., standard GMRES) for the preconditioned system* (2.24) *satisfies the bound*

$$\|\mathbf{r}^{m+\Delta m}\|^2 \le \left(1 - \frac{c_1^2}{c_2^2}\right)^m \|\mathbf{r}^0\|^2.$$

Proof. This uses the previous lemma with $\|\cdot\|_1 = \|\cdot\|_A$ and $\|\cdot\|_2 = \|\cdot\|$ (the Euclidean norm), together with inverse and norm equivalence estimates which imply (4.24) with $\underline{\gamma} = \underline{C}h^{d/2}$ and $\overline{\gamma} = \overline{C}\sqrt{a_{\max}}h^{d/2-1}$, where $\underline{C}$ and $\overline{C}$ are constants that are independent of all parameters. □

5 Numerical experiments

In this section we provide numerical results primarily for the two model problems: Firstly, an indefinite elliptic problem (where $\kappa > 0$)

$$-\mathrm{div}(A\nabla u) - \kappa u = f \qquad \text{in } \Omega, \tag{5.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{5.1b}$$

where the splitting in (2.4) used is given by

$$c^{+} = 0 \quad \text{and} \quad c^{-} = -\kappa. \tag{5.1c}$$

Secondly, a convection–diffusion problem

$$-\mathrm{div}(A\nabla u) + \mathbf{b} \cdot \nabla u = f \qquad \text{in } \Omega, \tag{5.2a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{5.2b}$$

where the splitting in (2.4) used is given by

$$c^{+} = 0 = c^{-}. \tag{5.2c}$$

Unless stated otherwise, in all our computations $\Omega = (0,1)^2$.

We first detail the common numerical setup. Our computations are performed using FreeFem [24] (`http://freefem.org/`), in particular within the `ffddm` framework. We perform experiments in 2D and discretise on uniform square grids with mesh spacing h, triangulating along one set of diagonals to form P1 elements. As the right-hand side f we impose a point source at the centre of the domain Ω.

To precondition the large sparse linear systems that arise from discretisation, unless otherwise stated, we utilise a two-level additive Schwarz preconditioner where the GenEO coarse space is incorporated additively, as in (2.23). This is used within a right-preconditioned GMRES iteration, which is terminated once a relative reduction of the residual of at least 10^{-6} has been achieved, or otherwise after a maximum of 1000 iterations (denoted in the tables below by '1000+'). For the local subdomains we use a uniform decomposition into N square subdomains which, unless stated otherwise, are then extended by adding only the fine-mesh elements which touch them (yielding a 'minimal overlap' scenario). Within the GenEO coarse space we take all eigenvectors corresponding to eigenvalues satisfying the eigenvalue threshold $\lambda < \lambda_{\max}$. Unless otherwise stated, we use the threshold parameter $\lambda_{\max} = 0.5$ which, in light of Definition 2.9, ensures that $\Theta \leq 2$ in our numerical experiments. Note that this means we do not make Θ particularly small in practice, although it is necessary in theory. To solve the local generalised eigenvalue problems we use ARPACK [27], as interfaced through FreeFem.

5.1 The case of homogeneous diffusion coefficient

To illustrate some of the basic properties of the GenEO method, we first consider some simple homogeneous test cases (i.e., with $A = I$), exploring how indefiniteness/non-self-adjointness affects performance. In Table 1 we display results for problem (5.1), varying the strength of the indefiniteness through increasing κ while fixing h (left) or varying $h \propto 1/\sqrt{\kappa}$ (right). As κ grows, the solution will be rich in the Laplace Dirichlet eigenmodes corresponding to eigenvalues near κ, and so this scaling is analogous to choosing a fixed number of points per wavelength in a Helmholtz problem. (In fact, here, approximately 100 points per wavelength). We see that the GenEO coarse space, built only from the positive (Laplacian) part of the problem (recall (5.1c)), works surprisingly well and provides a scalable solver for $\kappa \leq 1000$. However, once κ is too large the performance degrades and for $\kappa = 10000$ the method lacks robustness and scalability, the underlying problem now being highly indefinite. In comparing the left and right parts of Table 1, we see that choosing $h \propto (\sqrt{\kappa})^{-1}$ only mildly affects iteration counts for the range of κ in which the method works well.

Turning now to non-self-adjointness, Table 2 (left) gives results for (5.2) with $A = I$ and

$$\mathbf{b} = b\,(1 + \sin(2\pi(2y - x)))(2,1)^T, \tag{5.3}$$

for a fixed h. The non-self-adjointness can be increased by varying b. In this case, $\nabla \cdot \mathbf{b} = 0$ and so in (4.1) $C_2(\mathbf{b}, c^-) = C_1(\mathbf{b}, c^-)$, thus improving the estimate in Theorem 4.1. We observe

		N							N				
κ	h	4	16	36	64	100	κ	h	4	16	36	64	100
1	1/600	16	17	17	18	18	1	1/16	14	15	–	–	–
10	1/600	17	18	18	18	18	10	1/51	16	18	19	22	24
100	1/600	24	27	26	23	23	100	1/160	25	25	23	25	24
1000	1/600	40	98	102	113	89	1000	1/506	59	118	147	130	106
10000	1/600	144	431	660	1000+	1000+	10000	1/1600	161	446	821	1000+	1000+

Table 1: Iteration counts for the homogeneous, elliptic, indefinite problem (5.1), varying the parameter κ and the number of subdomains N for fixed $h = 1/600$ (left) and $h \propto 1/\sqrt{\kappa}$ (right).

	N						N				
b	4	16	36	64	100	b	4	16	36	64	100
1	18	18	18	18	18	1	18	18	18	18	18
10	28	26	23	22	21	10	28	26	23	22	21
100	35	34	30	28	27	100	39	43	35	29	25
1000	57	59	63	62	59	1000	72	107	71	74	63
10000	172	276	355	472	512	10000	201	450	708	835	837

Table 2: Iteration counts for the convection–diffusion problem (5.2), with homogeneous diffusion $A = I$ and convection coefficient (5.3) with zero divergence (left) and (5.4) with nonzero divergence (right), varying the constant b and the number of subdomains N for fixed $h = 1/600$.

that, also in this case, the GenEO coarse space works well and provides a scalable method for b up to at least 1000. For $b = 10000$ the convection term appears too strong and the iteration counts deteriorate.

We also consider problem (5.2) with a convection field with non-vanishing divergence, namely

$$\mathbf{b} = b\,(1 + \sin(2\pi(2x + y)))(2, 1)^T. \tag{5.4}$$

These results are given in Table 2 (right) and are similar to those in the left table, displaying robustness except in the most convection-dominated case. However, for higher b the iteration counts become larger compared to the zero-divergence case, illustrating that the strength of $\nabla \cdot \mathbf{b}$ does play a role, as predicted by the theory through the constant C_2 (given in (4.1)). We consider the impact of the convection field further in Section 5.4.

It is well-known that specialist discretisations of 'upwind' type, are required in order to obtain accurate solutions for practical meshes when the convection becomes large. However such approaches are typically not of the standard Galerkin form (2.11), and so an extension of the theory given here would be needed to treat these. Nevertheless since the standard Galerkin method treated here is effective if h is small enough relative to b, some sample results with $h \propto b^{-1}$ are given in Table 3. Here, for illustration, we compute only up to $b = 100$. Comparing the results from this refinement strategy to the fixed mesh results in Table 2, we see that the iteration counts have grown only mildly, and the GenEO approach remains very useful for moderate b.

We explore the dependence on mesh refinement further in Section 5.3; otherwise, in light of these results, we will typically consider h as fixed.

5.2 The case of a heterogeneous diffusion coefficient

We now consider the same experiments as in the previous subsection but with the diffusion coefficient A being heterogeneous. In particular, we will consider an "inclusions and channels"

		N				
b	h	4	16	36	64	100
1	1/16	14	15	–	–	–
10	1/160	23	21	21	20	20
100	1/1600	58	54	48	46	42

		N				
b	h	4	16	36	64	100
1	1/16	14	15	–	–	–
10	1/160	23	22	20	20	19
100	1/1600	59	69	56	46	40

Table 3: Iteration counts for the convection–diffusion problem (5.2), with homogeneous diffusion $A = I$ and convection coefficient (5.3) with zero divergence (left) and (5.4) with nonzero divergence (right), varying the constant b and the number of subdomains N for $h \propto 1/b$.

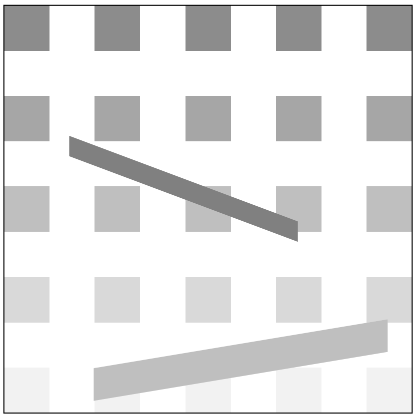


Figure 1: The heterogeneous function $a(\mathbf{x})$ within the inclusions and channels test problem. The shading gives the value of $a(\mathbf{x})$ with the darkest shade being $a(\mathbf{x}) = a_{\max}$, where $a_{\max} > 1$ is a parameter, and the background white taking the value $a(\mathbf{x}) = 1$.

problem, akin to that in [32]. Here, the heterogeneous coefficient is given by

$$A(\mathbf{x}) = a(\mathbf{x})I, \quad \text{where } a(\mathbf{x}) \text{ is given in Figure 1.} \tag{5.5}$$

The function $a(\mathbf{x})$ represents a number of inclusions, whose contrast with the background value of $a_{\min} = 1$ varies depending on the y-coordinate, along with two channels which run across the inclusions and link through a larger part of the domain, one with $a(\mathbf{x}) = a_{\max}$ and one with $a(\mathbf{x}) = \frac{1}{2}a_{\max}$. This represents a challenging heterogeneous problem for iterative solvers even in the self-adjoint and positive-definite case (i.e., $\mathbf{b} = 0$ or $\kappa = 0$). The strength of the heterogeneity is determined by the contrast parameter $a_{\max}$.

Results for problem (5.1) are given in Table 4 (left) while results for problem (5.2) with zero-divergence $\mathbf{b}$ are in Table 4 (right). Results for the case when $\nabla \cdot \mathbf{b} \neq 0$ are similar to those in the zero-divergence case, and thus omitted. We see that in all cases where the GenEO coarse space led to a robust method for the homogeneous problem, it does also for the heterogeneous problem and, moreover, is robust to the strength of the heterogeneity (here determined by $a_{\max}$). In the highly indefinite or highly non-self-adjoint cases, where the method struggles already for homogeneous problems, the lower iteration counts for some of the heterogeneous problems may stem from the fact that in some parts of the domain the heterogeneity in fact reduces the strength of the indefiniteness or of the convection term locally. For instance, consider the part Ω' of the domain where a is constant and has the largest value $a_{\max}$ (indicated by the darkest shade in Figure 1). Then, restricting (5.1) to Ω', we see that effectively (5.1) reduces to

$$-\Delta u - \left(a_{\max}^{-1}\kappa\right) u = a_{\max}^{-1} f,$$

hence reducing the effective indefiniteness in this region when $a_{\max} > 1$ is large.

		N				
κ	$a_{\max}$	4	16	36	64	100
1	5	18	18	18	18	18
1	10	18	18	18	18	19
1	50	18	18	19	19	19
10	5	18	18	18	19	19
10	10	18	18	19	19	19
10	50	18	18	20	19	19
100	5	27	30	26	23	22
100	10	26	26	24	23	22
100	50	23	22	22	20	20
1000	5	70	140	177	167	138
1000	10	58	118	146	131	108
1000	50	48	94	114	106	99
10000	5	193	524	906	1000+	1000+
10000	10	162	457	815	1000+	1000+
10000	50	114	318	704	888	1000+

		N				
b	$a_{\max}$	4	16	36	64	100
1	5	18	18	18	18	18
1	10	18	18	19	19	19
1	50	18	18	19	19	19
10	5	25	25	23	22	21
10	10	24	24	23	21	20
10	50	23	23	23	21	21
100	5	39	36	31	28	27
100	10	38	36	31	27	26
100	50	32	34	29	27	26
1000	5	51	59	57	57	54
1000	10	49	53	54	56	53
1000	50	53	57	55	54	51
10000	5	142	232	298	370	408
10000	10	121	180	261	320	328
10000	50	87	148	224	261	278

Table 4: Iteration counts for the heterogeneous, elliptic, indefinite problem (left) and for the convection–diffusion problem with zero-divergence convection coefficient (5.3) (right), varying the parameter κ (left) and the constant b (right), as well as $a_{\max} > 1$ and the number of subdomains N for fixed $h = 1/600$.

In Table 5 we display the size of the GenEO coarse space employed, along with comparison figures from the homogeneous case (labelled $a_{\max} = 1$). We note that, at most, only a small number of additional eigenvectors are required in the coarse space in order to treat the heterogeneity for any $a_{\max}$ used, thus showing robustness to contrasts in the diffusion coefficient. We also see that the size of the coarse space grows with N, however, the average number of eigenvectors per subdomain decreases with N so that we need to solve only small eigenvalue problems for a small number of eigenvectors. The size of the coarse space is also very small compared to the global problem size, here ranging from 0.06% to 0.5% as N increases from 4 to 100.

In conclusion, the GenEO method is well-equipped to deal with heterogeneities in the coefficient A, allowing us to treat the heterogeneous problem similarly to the homogeneous one.

	N				
$a_{\max}$	4	16	36	64	100
1	212 (53.0)	624 (39.0)	1060 (29.4)	1480 (23.1)	1800 (18.0)
10	209 (52.3)	613 (38.3)	1046 (29.1)	1447 (22.6)	1828 (18.3)
50	202 (50.5)	612 (38.3)	1055 (29.3)	1442 (22.5)	1845 (18.5)

Table 5: The size of the coarse space for the homogeneous problem (denoted $a_{\max} = 1$) and the heterogeneous inclusions and channels problem, varying $a_{\max}$ and the number of subdomains N for fixed $h = 1/600$. The average number of eigenvectors per subdomain, namely the size of the coarse space divided by N, is given in parenthesis. The total number of degrees of freedom for this problem is 361,201, yielding coarse space sizes ranging from 0.06% to 0.5% of the global problem size as N increases from 4 to 100.

	N			N		
h	16	36	100	16	36	100
$\frac{1}{200}$	87 (185)	115 (318)	90 (598)	87 (185)	115 (318)	90 (598)
$\frac{1}{400}$	89 (419)	107 (661)	91 (1214)	88 (225)	109 (356)	94 (639)
$\frac{1}{600}$	94 (612)	114 (1055)	99 (1845)	92 (227)	116 (384)	99 (672)
$\frac{1}{800}$	95 (847)	115 (1329)	100 (2478)	93 (228)	116 (379)	100 (680)
$\frac{1}{200}$	0.46%	0.79%	1.48%	0.46%	0.79%	1.48%
$\frac{1}{400}$	0.26%	0.41%	0.75%	0.14%	0.22%	0.40%
$\frac{1}{600}$	0.17%	0.29%	0.51%	0.06%	0.11%	0.18%
$\frac{1}{800}$	0.13%	0.21%	0.39%	0.04%	0.06%	0.11%

Table 6: Above: Iteration count and size of the coarse space (in brackets) for minimal overlap $\delta \sim h$ (left) and for generous overlap $\delta \sim H$ (right) for problem (5.1) with heterogeneous A as in (5.5), varying h and N for $\kappa = 1000$ and $a_{\max} = 50$ fixed. Below: The size of the coarse space as a percentage of the global problem size.

5.3 Dependence on mesh refinement

In this subsection we explore dependence upon the mesh spacing h. We consider problem (5.1) with A given as in (5.5) and we set $\kappa = 1000$ and $a_{\max} = 50$. Table 6 (left) shows how the iteration count and the size of the coarse space vary with h. We note that the iteration count remains stable as h decreases and the number N of subdomains grows. On the other hand, the size of the coarse space appears to grow at a rate slightly slower than $\mathcal{O}(h^{-1})$ as h is decreased. This is in full agreement with similar results in [3, Table 1] for positive-definite problems in the case of minimal overlap $\delta \sim h$.

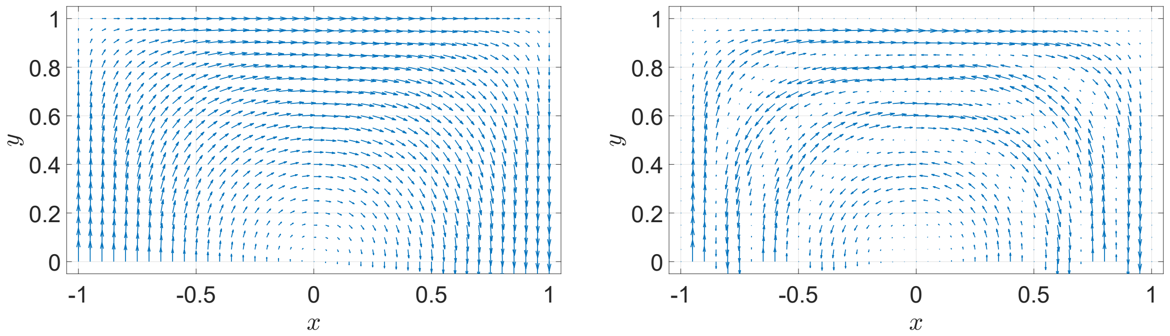
As in [3], we repeat our experiment for the case of generous overlap $\delta \sim H$. We fix the subdomain size H and the overlap δ independent of h, in particular to that given by minimal overlap on the coarsest mesh size $h = \frac{1}{200}$, such that $\delta = \frac{1}{100}$. The partition of unity is chosen to vary smoothly over the entire extent of the overlap. (For details of the implementation within FreeFem, see [26, Section 4.1].) The iteration counts presented in Table 6 (right) for this setup remain stable and are essentially identical to those for minimal overlap in the left table, but in addition, we now also observe that the size of the coarse space remains bounded as $h \to 0$, which is again in agreement with the results in [3, Table 1].

In all cases, our results show that the dimension of the coarse space depends on N, and grows approximately as $\mathcal{O}(N^{2/3})$; this means the number of eigenvectors chosen per subdomain decreases with N at approximately the rate $\mathcal{O}(N^{-1/3})$. Further, Table 6 (below) provides the size of the coarse space as a percentage of the global problem size and we note that this is both small (at most 1.48%) and decreases with h.

5.4 The effect of the convection field

For problem (5.2) the specific form of $\mathbf{b}$ can play an important role. Here we consider some further examples to illustrate this point and show how the divergence of $\mathbf{b}$ can dictate performance, as predicted by the theory. We start with the homogeneous case $A = I$, and with a circulating convection field. Here the problem is posed on $(-1, 1) \times (0, 1)$ and the convection coefficient is given as

$$\mathbf{b} = b\,(2y(1 - x^2), -2x(1 - y^2))^T, \tag{5.6}$$



(a) Without oscillations: $\mathbf{b}$ given by (5.6). (b) With oscillations: $\mathbf{b}$ given by (5.7) with $n = 4$.

Figure 2: Circulating convection fields with and without oscillation in the radial direction.

b	N = 4	16	36	64	100
1	18	18	18	18	18
10	24	24	22	21	20
100	43	44	41	36	33
1000	31	34	35	35	36
10000	143	231	282	314	339
–	316	931	1584	2170	2796

n	N = 4	16	36	64	100
1	30	33	35	35	36
2	35	44	49	45	49
4	44	64	65	57	57
8	46	66	70	65	60

Table 7: Iteration counts for problem (5.2) with $A = I$ and with circulating convection coefficient for $h = 1/600$. Left: Using (5.6) and varying the parameter b and the number of subdomains N; the coarse space size (independent of b) is tabulated in the final row. Right: Using (5.7) with radial oscillations and $b = 1000$, varying the parameter n and the number of subdomains N.

illustrated in Figure 2(a). Results varying b, which controls the strength of the convective term, are given in Table 7 (left). As before, GenEO behaves robustly for b up to at least 1000.

As well as varying the strength of the convective term, we can also add local variation. To this end, we add sinusoidal variation of increasing frequency in the radial direction (perpendicular to the characteristics of the convection field). That is, we consider

$$\mathbf{b} = b\sin(n\pi(1-x^2)(1-y^2))(2y(1-x^2), -2x(1-y^2))^T, \tag{5.7}$$

for a parameter n; see Figure 2(b). Note that for any choice of n this convection coefficient is divergence-free since $\nabla \cdot \mathbf{b} = 0$. Thus, here we illustrate the effect of local variation without changing the divergence. Results for $b = 1000$ and varying n are given in Table 7 (right), where we see that iteration counts slightly increase with the frequency parameter n but that this remains mild and the coarse space continues to provide a robust preconditioner as N increases.

We now explore the role the local variation of the divergence of $\mathbf{b}$ plays by considering a simple unidirectional convection field with sinusoidal variation added in such a way as to vary $\nabla \cdot \mathbf{b}$. In particular, we consider

$$\mathbf{b} = b\,(1+\sin(m\pi(2x+y)))(1+\sin(2\pi(2y-x)))(2,1)^T, \tag{5.8}$$

see Figure 3, on the unit square. In this case, we also choose the heterogeneous diffusion coefficient A in (5.5) to incorporate the full range of heterogeneity.

Here $\mathbf{b}$ includes oscillation both across characteristics and along characteristics. The term $(1+\sin(2\pi(2y-x)))$ introduces oscillations that do not change the strength of the divergence. On the other hand, the term $(1+\sin(m\pi(2x+y)))$ introduces oscillations along characteristics

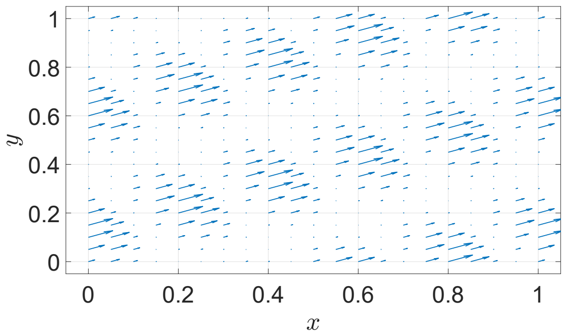


Figure 3: Unidirectional convection field with oscillations: $\mathbf{b}$ given by (5.8) with $m = 4$.

	N				
m	4	16	36	64	100
0	53	57	55	54	51
2	56	85	88	95	78
4	50	97	143	162	123
8	56	79	171	133	181

Table 8: Iteration counts for the heterogeneous convection–diffusion problem with oscillating unidirectional convection coefficient (5.8), varying the parameter m (which controls the size of $\nabla \cdot \mathbf{b}$) and the number of subdomains N, with fixed $h = 1/600$, $b = 1000$ and $a_{\max} = 50$.

and as we increase their frequency through the parameter m we increase the strength of the divergence locally. This can be seen visually in Figure 3 where we have $m = 4$ oscillations along any unit vector parallel to $(2, 1)^T$ and 2 oscillations along any unit vector perpendicular to $(2, 1)^T$. To see how this affects performance we fix $b = 1000$ and $a_{\max} = 50$ and display results for increasing m in Table 8. This time we see that increasing the oscillations has a significant impact on the iteration counts and the GenEO coarse space starts to struggle if m gets too big. This observation aligns with the theory which suggests that a large divergence can hamper the effectiveness of the coarse space.

To summarise, the GenEO coarse space is able to handle variations in the convection field to an extent. However, should $\nabla \cdot \mathbf{b}$ become too large then the approach can lose robustness.

5.5 Time evolution for convection–diffusion problems

Finally, we also include one experiment on time evolution for convection–diffusion problems. The underlying time-dependent problem for $u(\mathbf{x}, t)$ over the time-interval $(0, T)$ is given by

$$\frac{\partial u}{\partial t} - \operatorname{div}(A\nabla u) + \mathbf{b} \cdot \nabla u + \kappa u = g \qquad \text{in } \Omega \times (0, T), \tag{5.9a}$$

$$u = 0 \qquad \text{on } \partial\Omega \times (0, T), \tag{5.9b}$$

$$u = u_0 \qquad \text{on } \Omega \times \{0\}\,. \tag{5.9c}$$

Consider using an implicit scheme in time to yield a semi-discrete problem. In particular, by way of example, we consider the backward Euler scheme with time-step $\Delta t > 0$ and set (without loss of generality) $\kappa = 0$, yielding

$$-\operatorname{div}(A\nabla u^n) + \mathbf{b} \cdot \nabla u^n + \frac{1}{\Delta t} u^n = g^n + \frac{1}{\Delta t} u^{n-1} \qquad \text{in } \Omega, \tag{5.10a}$$

$$u^n = 0 \qquad \text{on } \partial\Omega, \tag{5.10b}$$

Δt	N 4	16	36	64	100
1000	53/53/44	57/56/49	55/55/51	54/54/51	51/51/49
0.1	53/53/44	57/56/49	55/55/51	54/54/51	51/51/49
0.001	50/50/43	53/53/48	53/53/50	52/52/50	50/50/48

Table 9: Iteration counts with $\Delta t_0 = 10$ (left), 0.1 (middle), and 0.001 (right) for one time-step (5.11) of backward Euler applied to the time-dependent convection–diffusion problem with heterogeneous diffusion coefficient (5.5) and unidirectional convection coefficient (5.3), varying the time-step Δt and the number of subdomains N, for fixed $h = 1/600$, $b = 1000$ and $a_{\max} = 50$.

where $u^n(\mathbf{x}) \approx u(\mathbf{x}, n\Delta t)$ and $u^0 = u_0$ with a similar convention for g. The canonical expression of the semi-discrete problem at each time-step takes the form

$$-\mathrm{div}(A\nabla u) + \mathbf{b} \cdot \nabla u + \frac{1}{\Delta t} u = f \qquad \text{in } \Omega, \tag{5.11a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{5.11b}$$

for some f and it is this generic problem we consider solving. Other implicit time-stepping schemes will lead to similar PDE problems at each time-step, so this investigation covers the general case.

Since $1/\Delta t > 0$, it could be put completely into the coefficient c^+, as defined in (2.4), and thus used in the definition of the GenEO coarse space. However, to allow better efficiency in the case of variable time-stepping it would be better to build a preconditioner that is independent of the time-step size. So we suppose that a 'median time-step' Δt_0 is chosen and we perform the splitting (2.4) with

$$c^+ = 1/\Delta t_0 \quad \text{and} \quad c^- = 1/\Delta t - 1/\Delta t_0.$$

With this choice, the GenEO preconditioner is built—once only, before time-stepping takes place—incorporating the $1/\Delta t_0$ term, i.e., using the bilinear form

$$a_{\Omega_i}(u, v) = \int_{\Omega_i} \left(A\nabla u \cdot \nabla v + \frac{1}{\Delta t_0} uv \right) dx.$$

We study the utility of this choice by varying parameters Δt_0 and Δt.

We take $\Omega = (0, 1)^2$, A as in (5.5), and consider the divergence-free convection coefficient (5.3), fixing $b = 1000$ and $a_{\max} = 50$. Results are given in Table 9 where we see that the GenEO preconditioner is robust to changes in both Δt_0 and Δt (with either $\Delta t \leq \Delta t_0$ or $\Delta t > \Delta t_0$), thus providing an effective and efficient preconditioner for implicit time-stepping of non-self adjoint or indefinite time-dependent problems.

5.6 Alternative two-level preconditioners and GenEO-type coarse spaces

The theory in this paper and the numerical results above are for the classical, additive two-level Schwarz preconditioner with the GenEO coarse space in (2.32). The underlying generalised eigen-problem in (2.31) is based on the self-adjoint and positive-definite subproblem (2.5), which inevitably ceases to provide a robust preconditioner when the indefiniteness or non-self-adjointness is very strong.

Especially for indefinite problems, such as (5.1), the *Restricted Additive Schwarz* (RAS) preconditioner [8] along with a deflation approach to incorporate the coarse space has previously been shown to lead to significantly lower iteration counts and better robustness with respect to the indefiniteness [5, 6].

		N							
κ	$a_{\max}$	16	36	64	100	16	36	64	100
10	5	18	18	19	19	9	9	9	9
10	10	18	19	19	19	9	9	9	9
10	50	18	20	19	19	9	9	9	9
100	5	30	26	23	22	11	10	10	9
100	10	26	24	23	22	11	10	10	9
100	50	22	22	20	20	9	9	9	9
1000	5	140	177	167	138	62	76	70	57
1000	10	118	146	131	108	50	62	55	45
1000	50	94	114	106	99	38	46	43	39
10000	5	524	906	1000+	1000+	297	504	628	753
10000	10	457	815	1000+	1000+	234	445	548	644
10000	50	318	704	888	1000+	171	372	464	538

Table 10: Iteration counts for problem (5.1) with heterogeneous A as in (5.5), using the GenEO coarse space based on (2.31) and varying κ, $a_{\max} > 1$ and N with $h = 1/600$ fixed. Left: Classical two-level, additive Schwarz. Right: Restricted additive Schwarz and deflation.

For the remainder of this section we return to problem (5.1), and start by repeating the experiments reported in Table 4 (left), but this time we use RAS instead of additive Schwarz for the one-level part of the preconditioner, and we combine this with the coarse solve using a deflation approach. Otherwise the setup is identical to that in Table 4 (left). See, e.g., [5] for more details of this 'RAS plus deflation' algorithm. Results for this 'RAS plus deflation' approach are given in Table 10 (right) which can be compared with the corresponding results from Table 4 (given again here on the left for convenience). We see that the 'RAS plus deflation' algorithm requires roughly half the number of iterations needed by the standard algorithm, but it does not yield any significantly improved robustness with respect to the indefiniteness.

In [5] we also compared the GenEO coarse space based on the eigenproblem (2.31), denoted Δ-GenEO in [5], with an alternative approach first introduced for the Helmholtz problem in [6], going under the name of $\mathcal{H}$-GenEO.

Definition 5.1. *($\mathcal{H}$-GenEO eigenproblem for problem (5.1)). For each $j = 1, \ldots, N$, we define the following generalized eigenvalue problem*

$$\text{Find} \quad (p, \lambda) \in \widetilde{V}_j \backslash \{0\} \times \mathbb{R} \quad \text{such that} \quad b_{\Omega_j}(p, v) = \lambda\, a_{\Omega_j}(\Xi_j(p), \Xi_j(v)), \quad \text{for all } v \in \widetilde{V}_j, \tag{5.12}$$

where $b_{\Omega_j}(\cdot,\cdot)$ is the bilinear form corresponding to the indefinite operator $-\Delta - \kappa$ and Ξ_j is the local partition of unity operator from Definition 2.7.

Recalling (2.31), we see that the key difference in $\mathcal{H}$-GenEO compared with Δ-GenEO is the use of the bilinear form $b_{\Omega_j}(\cdot,\cdot)$ of the underlying indefinite problem on the left-hand side. Note that this also implies that the eigenvalues λ in (5.12) can be negative. Thus, when incorporating eigenvectors into the coarse space using the threshold $\lambda < \lambda_{\max}$ we also include all negative eigenvalues (see again [5] for details).

In our final set of results, Table 11, we repeat the experiments in Table 10 using the GenEO coarse space based on the $\mathcal{H}$-GenEO eigenproblem in (5.12), again with an otherwise identical setup. We see that for small κ the iteration counts are almost identical. For large κ, we see that $\mathcal{H}$-GenEO leads to much lower iteration counts, such that GMRES converges in less than 50 iterations even for $\kappa = 10000$ when using the 'RAS plus deflation' approach. Unfortunately, a theoretical justification for this excellent performance of the $\mathcal{H}$-GenEO method is still lacking. We note that for larger κ the coarse space induced by the $\mathcal{H}$-GenEO eigenproblem in (5.12) is

		N							
κ	$a_{\max}$	16	36	64	100	16	36	64	100
10	5	18	18	19	19	9	9	9	9
10	10	18	19	19	19	9	9	9	9
10	50	18	19	19	19	9	9	9	9
100	5	18	19	19	19	9	9	9	9
100	10	18	19	19	19	9	9	9	9
100	50	18	19	19	19	8	9	9	9
1000	5	43	36	29	32	12	12	11	12
1000	10	36	43	23	24	11	11	10	10
1000	50	28	31	24	26	12	11	11	12
10000	5	134	173	192	258	35	35	43	45
10000	10	113	147	191	192	21	29	41	34
10000	50	92	109	173	182	19	25	48	26

Table 11: Iteration counts for problem (5.1) with heterogeneous A as in (5.5), using the $\mathcal{H}$-GenEO coarse space based on (5.12) and varying κ, $a_{\max} > 1$ and N with $h = 1/600$ fixed. Left: Classical two-level, additive Schwarz. Right: Restricted additive Schwarz and deflation.

slightly larger than that induced by the eigenproblem in (2.31) (see [5, Table 2]). Nonetheless, this increase is relatively modest compared to the reduction in iteration counts that can be achieved.

These results provide an insight into the power of using generalised eigenproblems to yield coarse spaces for challenging indefinite or non-self-adjoint problems, which has yet to be fully explored even numerically. A number of theoretical obstacles remain in order to analyse the effect of such GenEO-type methods, but theory in this direction may provide a greater understanding of where such approaches can be successful and how to define appropriate, problem-dependent generalised eigenproblems to incorporate the salient properties of the underlying problem.

References

[1] E. Agullo, L. Giraud, and L. Poirel, *Robust preconditioners via generalized eigenproblems for hybrid sparse linear solvers*, SIAM J. Matrix Anal. Appl. 40, 417–439, 2019.

[2] I. Babuska and R. Lipton, *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul. 9, 373–406, 2011.

[3] P. Bastian, R. Scheichl, L. Seelinger and A. Strehlow, *Multilevel spectral domain decomposition*, to appear in SIAM J. Sci. Comput., `https://arxiv.org/abs/2106.06404`, 2021.

[4] N. Bootland and V. Dolean, *On the Dirichlet-to-Neumann coarse space for solving the Helmholtz problem using domain decomposition.* In: Numerical Mathematics and Advanced Applications ENUMATH 2019, *Lect. Notes Comput. Sci. Eng.*, vol. 139, pp. 175–184. Springer, Cham, 2021.

[5] N. Bootland, V. Dolean, I. G. Graham, C. Ma and R. Scheichl, *GenEO coarse spaces for heterogeneous indefinite elliptic problems*, to appear in Proc. 26th Int. Domain Decomposition Conference (2020), `https://arxiv.org/abs/2103.16703`, 2021.

[6] N. Bootland, V. Dolean, P. Jolivet and P.-H. Tournier, *A comparison of coarse spaces for Helmholtz problems in the high frequency regime*, Comput. Math. Appl. 98, 239–253, 2021.

[7] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer Science & Business Media, 2007.

[8] X.C. Cai and M. Sarkis, *A restricted additive Schwarz preconditioner for general sparse linear systems*, SIAM J. Sci. Comput. 21, 239–247, 1999.

[9] X.C. Cai and O.B. Widlund, *Domain decomposition algorithms for indefinite elliptic problems*, SIAM J. Sci. Stat. Comput. 13(1), 243–258, 1992.

[10] X.C. Cai and J. Zou, *Some observations on the l_2 convergence of the additive Schwarz preconditioned GMRES method*, Numer. Linear Algebra Appl. 9, 379–397, 2002.

[11] L. Conen, V. Dolean, R. Krause and F. Nataf, *A coarse space for heterogeneous Helmholtz problems based on the Dirichlet-to-Neumann operator*, J. Comput. Appl. Math. 271, 83–99, 2014.

[12] V. Dolean, P. Jolivet, and F. Nataf, *An Introduction to Domain Decomposition Methods: Algorithms, Theory and Parallel Implementation*, SIAM, Philadelphia, 2015.

[13] Y. Efendiev, J. Galvis and T.-Y. Hou, *Generalized multiscale finite element methods (GMsFEM)*, J. Comput. Phys. 251, 116–135, 2013.

[14] S.C. Eisenstat, H.C. Elman, and M.H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal. 20, 345–357, 1983.

[15] J. Galvis, E.T. Chung, Y. Efendiev and W.T. Leung, *On Overlapping Domain Decomposition Methods for High-Contrast Multiscale Problems*. In: P. Bjørstad et al. (eds), Domain Decomposition Methods in Science and Engineering XXIV, *Lect. Notes Comput. Sci. Eng.*, vol 125, pp. 45–57. Springer, Cham, 2018.

[16] J. Galvis and Y. Efendiev, *Domain decomposition preconditioners for multiscale flows in high-contrast media*, Multiscale Model. Simul. 8, 1461–1483, 2010.

[17] J. Galvis and Y. Efendiev, *Domain decomposition preconditioners for multiscale flows in high contrast media: reduced dimension coarse spaces*, Multiscale Model. Simul. 8, 1621–1644, 2010.

[18] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, 2001.

[19] S. Gong, I.G. Graham and E.A. Spence, *Domain decomposition preconditioners for high-order discretisations of the heterogeneous Helmholtz equation*, IMA J. Numer. Anal. 41, 2139–2185, 2021.

[20] J. Gopalakrishnan and J. Pasciak, *Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations*, Math. Comput. 72(241), 1–15, 2001.

[21] I.G. Graham, P.O. Lechner, and R. Scheichl, *Domain decomposition for multiscale PDEs*, Numer. Math. 106, 589–626, 2007.

[22] I.G. Graham, E.A. Spence and E. Vainikko, *Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption*, Math. Comput. 86, 2089-2127, 2017.

[23] W. Hackbusch, *Elliptic Differential Equations Theory and Numerical Treatment*, Second Edition, Springer, 2017.

[24] F. Hecht, *New development in FreeFem++*, J. Numer. Math., 20(3–4), 251–266, 2012.

[25] A. Heinlein, A. Klawonn, J. Knepper, and O. Rheinbach, *Adaptive GDSW coarse spaces for overlapping Schwarz methods in three dimensions*, SIAM J. Sci. Comput. 41, A3045–A3072, 2019.

[26] P. Jolivet, V. Dolean, F. Hecht, F. Nataf, C. Prud'Homme, and N. Spillane, *High performance domain decomposition methods on massively parallel architectures with FreeFEM++*, J. Numer. Math., 20(3–4), 287–302, 2012.

[27] R. Lehoucq, D. Sorensen, and C. Yang, *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, SIAM, Philadelphia, 1998.

[28] G. Leoni, *A First Course in Sobolev Spaces*, American Mathematical Soc., 2017.

[29] C. Ma, R. Scheichl and T. Dodwell, *Novel design and analysis of generalized FE methods based on locally optimal spectral approximations*, https://arxiv.org/abs/2103.09545, 2021.

[30] C. Ma and R. Scheichl, *Error estimates for fully discrete generalized FEMs with locally optimal spectral approximations*, https://arxiv.org/abs/2107.09988, 2021.

[31] F. Nataf and P.-H. Tournier, *A GenEO domain decomposition method for saddle point problems*, https://hal.archives-ouvertes.fr/hal-02343808, 2021.

[32] F. Nataf, H. Xiang, V. Dolean, and N. Spillane, *A coarse space construction based on local Dirichlet-to-Neumann maps*, SIAM J. Sci. Comput., 33(4), 1623–1642, 2011.

[33] A.H. Schatz, *An observation concerning Ritz–Galerkin methods with indefinite bilinear forms*, Math. Comput., 28(128), 959–962, 1974.

[34] A.H. Schatz and J.P. Wang, *Some new error estimates for Ritz–Galerkin methods with minimal regularity assumptions*, Math. Comput., 213(65), 19–27, 1996.

[35] B.F. Smith, P. Bjorstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, 1996.

[36] N. Spillane, *Toward a new fully algebraic preconditioner for symmetric positive definite problems*, https://hal.archives-ouvertes.fr/hal-03187092, 2021.

[37] N. Spillane, F. Nataf, V. Dolean, P. Hauret, C. Pechstein, and R. Scheichl, *Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps*, Numer. Math. 4(126), 741–740, 2014.

[38] A. Toselli and O. Widlund, *Domain Decomposition Methods: Algorithms and Theory*, Springer Series in Computational Mathematics. Springer-Verlag, 2005.